UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

WORKING PAPERS
No. 6/2023 (413)

MODELLING SUBJECTIVE ATTRACTIVENESS

KONRAD LEWSZYK
PIOTR WÓJCIK

WARSAW 2023

WORKING PAPERS 6/2023 (413)

Modelling Subjective Attractiveness

Konrad Lewszyk, Piotr Wójcik

University of Warsaw, Faculty of Economic Sciences and Data Science Lab WNE UW
Corresponding author: pwojcik@wne.uw.edu.pl

Abstract: Attractive people obtain greater economic and reproductive success. This article attempts to grasp individual preferences of facial attractiveness and create reliable models that will accurately predict a beauty score on a binary and quintary scale. Based on extensive research conducted on factors of attractiveness, we derive the most important facial features that have the highest impact in beauty perception. Based on a sample of 681 images of faces using facial a landmark detector. We derive various numerical features represented by face characteristics and. The application of various machine learning algorithms shows that attractiveness can be predicted accurately based on facial characteristics. In addition, we show that indeed the attractiveness is subjective as the same features have different importance for different subjects.

Keywords: Attractiveness, beauty-premium, image processing, machine learning, predictive models

JEL codes: C40, C53, J71

1. Introduction

Attractiveness can be understood as a degree to which a person's physical characteristics are considered aesthetically pleasing to a particular person or to a group of people. The evolutionary task of attractiveness is reproduction. The more attractive a male/female appears the more partners she or he will attract. Not only does attractiveness drive reproduction, it also reinsures that the offspring will be fit and healthy. It turns out that attractiveness is of crucial importance in sexual reproduction, since it determines directly the benefits passed on to the offspring (Jokela 2009, Prokop and Fedor 2011, Pflüger et al. 2012). On the other hand attractiveness plays an additional role in economic success in modern times. Good looks have been linked to better chance of employment (Pfeifer 2012, Stinebrickner et al. 2019) and have a higher probability of earning more (so called *beauty premium* – Kanazawa and Still 2019, Dossinger et al. 2019, Abueg et al. 2020). In both dimensions of success (reproductive and economic) the researchers point at the facial attractiveness (Luxen and Van De Vijver 2006, Scholz and Sicinski 2015) as the main indicator of the overall attractiveness.

The correlation between attractiveness and health is strong and the perception of beauty is dependent on the culture, the fact is that attractiveness is still largely a subjective matter with many disagreeing on relative attractiveness of faces, which means that the matter of attractiveness is left to the eye of the beholder. Numerous studies have found that our subjective judgement can be influenced by the personality type we seek others. Humans tend to judge higher faces which to them resemble a personality trait they are looking for in their mates (Little, 2006). It is also known that our own looks alter our own perception of beauty. Faces that are closer in resemblance to our own appear to be more attractive (DeBruine, 2004). Finally, in a study where affiliated members (spouses, siblings, close friends) were asked to rate given faces on attractiveness based on their judgement, their ratings were in a significantly greater agreement than when compared to ratings of strangers (Bronstad, 2007). The matter of subjectiveness has to be understood better in order to address the low self esteem issues and the matter of unequal career opportunities related to looks.

The **goal of this article** is to attempt to capture attractiveness models of individual people assessing the attractiveness of women faces (presented on images) and identify the physical characteristics of faces that drive attractiveness preferences. Our **main research hypothesis** claims that based on the characteristics of a face one can accurately predict the subjective attractiveness of the person. The **supplementary hypothesis** assumes that the attractiveness assessment is

subjective, i.e. different characteristics of faces influence the attractiveness score for different subjects. We collected 681 images of women faces and asked 5 experiment participants to rate all faces on two scales – binary and on a scale from 1 to 5 (quintary). Using a specific image processing tool (facial landmark detection) we identified 68 different points of each face and based on that calculated numerical characteristics related to the shape of face elements (eye, eyebrow, nose, lips, ears, etc.). These characteristics were used as predictors of subjective attractiveness. The data was randomly divided into train and test sets and we applied xgboost classification algorithm for each of the participants. Therefore, ten models were applied in total (5 for binary classification and 5 for quintary classification). The models were evaluated with four metrics (accuracy, precision, recall and f1). We tried to understand who a particular person finds or does not find attractive based on facial characteristics.

The remaining part of the article is structured in the following way. In section 1 we describe the supporting literature for our study. We rely on literature that describes the importance of looks in economy, and we also review studies that reveal which facial characteristics are most influential in perception of beauty. Section 2 explains the methodology we use in our study. We first obtain the pictures of faces, then proceed to the rating process completed by participants of our study and finally we describe the methods for modeling and predicting attractiveness based on the preferences of our participants. Section 3 is devoted to the description of the facial dataset and conversion of each face to numerical characteristics that best describe that particular face. Section 4 includes the empirical part where we evaluate our models and see if the predictions are accurate. The article ends with the summary of conclusions.

2. Literature review

2.1. Attractiveness in life and economy

One of the most sought out characteristics across different cultures and species (such as macaques, zebra finches or barn swallows) is symmetry (Thornhill, 1993). In ideal conditions a face should grow up to be perfectly symmetrical. Nonetheless no one grows up in ideal conditions and various aspects interfere with our development, such as poor diet, malnutrition, infectious diseases, illnesses, lack of sleep or exercise. All these factors influence our facial symmetry. Knowing this, an ideal facial symmetry is probably impossible to find, so healthiest individuals would be those

deviating as little as possible from the supposed ideal symmetry. Therefore it has been proposed that the level of symmetry represents developmental stability and is an indicator of how well an individual was able to fight off these interruptions and how strong his immune system is (Moller, 1990).

Another important aspect of attractiveness is averageness. When a large number of faces is blended together, one can derive an average from the most important characteristics, which are skin color, eye color, but also coordinates of the facial features, such as eyes, nose, eyebrows, mouth and others. Previous studies have found that these blended faces that represent the average characteristics of a population rate higher in attractiveness than those individual faces from which they are made (Rhodes, 2001). It has been suggested that participants rated the average faces higher than those that deviate from the average because they avoided distinctive features. The deviations in their facial features might suggest they are suffering from an illness or carry unhealthy genes (Rhodes, 1996), while the average faces are known to poses more heterozygosity in their DNA known as the major histocompatibility complex (MHC). This means that the DNA of people with average faces has more varied genes for immune functions and is able to produce proteins to fight off a broad range of pathogens (Thornhill, 1993).

The connection between attractiveness and health carries over to aspect of body shape and the BMI index as well. It has been deducted that in developed societies men think women who are in the middle of the healthy body mass index range (BMI between 19 and 24.5) look the healthiest, while women who are on the lower end of that spectrum are deemed to be most attractive (Tovee, 1998). Naturally the weight of an individual is reflected in the face and referred to as facial adiposity (apparent weight in the face). Facial adiposity consistent with the same body weight and BMI between 19 and 24.5 is also found to look the healthiest and most attractive (Coetzee, 2011). Connection between health and body weight is common knowledge to all. Overweight and obese individuals are at a higher risk of developing various diseases such as diabetes, heart diseases, strokes or cancers. Increased levels of facial adiposity have been directly linked to higher blood pressure and lower immunity to infections (Coetzee, 2009).

The previously discussed examples of connections between health and perceived attractiveness hint that perceived health is directly linked with perceived attractiveness and should be universal across cultures and equal. Interestingly our perception of beauty and health can be strongly shaped by our culture. In parts of the world where food is more scarce and valuable, larger bodies represent

the ability to obtain more nutrition. In sub-Saharan Africa and Malaysia attractiveness based preferences shift towards individuals with higher BMI index (Tovee, 2005).

But our preferences can also be shaped by our culture, not just the environment in which we live. In countries with higher fertility rates and ubiquitously expressed preference to have sons over daughters men tend to look for females with wider hips. It has been found that wider hips in females signify a higher fertility potential. A study conducted in Jamaica found a positive and significant relationship between waist circumference and the number of sons a particular woman had (Yu, 1999). Similar studies have been conducted in Texas and England and both have drawn same conclusions (Manning, 1996). Another example of cross-cultural factor in the judgement of attractiveness is economy. Women who live in countries with higher income inequality and competition have a strong preference for more masculine looking men, who might be better suited for competing in such environments (Brooks, 2011).

Physical attractiveness is strongly related to psychological well being and lower self esteem connected to feeling of unattractiveness has shown to be one of causes of distress and depression (Gupta, 2016). A recent survey conducted by Dove revealed that 96% of women would not use the word ‘beautiful’ to describe themselves and 76% of female respondents admitted to not feeling attractive (Nithya, 2015).

Numerous studies have explored the link between attractiveness and reproductive success in today’s times. How a woman looks in her youth strongly determines her potential for marriage and reproductive success. Women who show more desirable features in their youth are correlated with a higher number of offspring later in life (Plfuger, 2012). Interestingly it has been found that highly attractive females were more likely not only to become parents, but also choose to have a second child when in comparison with their less attractive counterparts. Nonetheless, attractive women are less likely to have a third or a fourth child than less attractive women. On the other hand male attractiveness was positively linked with having one, two, three or even four children. Finally attractive individuals had a higher probability of getting married in their youth (Jokela, 2009).

While the aspect of attractiveness as a driving force of reproduction is worldwide, the role of attractiveness in economy is also ubiquitous. Pfeifer (2011) conducted a study in Germany to investigate the link between looks and education, job prospects and employment probabilities. The study was based on a series of interviews, where the interviewer first rates the respondent, conducts the interview and then completes the rating for the second time at the end of the interview to see if

perception of attractiveness has been altered by the behavior of the interviewee. The participants of the study also rated themselves. The rating data was combined with personal data of the respondents. The personal data included wages, employment, education, political behavior etc. The conclusions of the study showed with significance that more attractive people are more likely to be employed and earn on average higher wages. The same conclusions have been drawn by studies conducted in Canada, China and USA (Hamermesh, 1994; 2002; French, 2002, Fletcher, 2009). All of the economies described belong to the first world, and China and USA are the two biggest economies in the world. This means that collectively the four economies described in this paragraph give an overall worldwide view on importance of attractiveness in employment and career success.

Just like Anýžová and Matějů (2018), Pfeifer disputes the idea that cognitive aspects play a significant role in comparison to looks solely. Both studies debated whether the lower wages and career prospects aren't necessarily connected solely to looks. A person who would score lower on attractiveness scale might suffer from lower self esteem and lower confidence. This would explain that the phenomenon of higher pay given to attractive people is a joint effect of good looks and high self esteem. The conclusions state that the effect of low self esteem is not significant and that looks play a major role, which means there is no joint effect in the good looks and labor marker phenomenon.

Scholz and Sicinski (2015) explored the long term effect of attractiveness on lifetime earnings and used longitudinal data on male high school graduates and their subsequent job market earnings throughout their life in their mid 30s and mid 50s. In addition to earnings, IQ, high school activities, measures of confidence and other characteristics were also included in the dataset. What they found was a durable and persistent and strong correlation between facial attractiveness of men and earnings in mid 30s and 50s. The IQ and the included extensive set of characteristics did not play a significant role, and the "beauty premium" (a term that describes the economical advantage for highly attractive people) effect was present regardless of educational attainment, household characteristics of occupational choices.

One of the reasons why attractive people can achieve more success in their careers is the stereotype known as "What is beautiful is good" (Dion, 1972). This hypothesis states that people who possess more attractive traits are related to perception of desirable interpersonal traits.

Furthermore attractive people are seen as more motivated to form social bonds. The stereotype has been found to be present across different cultures and has been confirmed by numerous studies since the first study on the topic in 1972. A recent study exploring possibilities on how to tackle beauty bias in the hiring process found that highly attractive individuals earn roughly 20% more and are recommended more frequently for promotions. Additionally attractive individuals can often enjoy their privilege without a justification in their economic productivity (Nault, 2020).

Nonetheless recent research shows that the "what is good is beautiful" stereotype is not as straightforward as it seems and is not entirely left to beauty alone. An analysis of college graduates and their careers has shown that attractiveness played a significant role in the hiring process and advancement in the career in jobs that required substantial amounts of interpersonal interactions. On the other hand jobs that dealt with information strictly and relied less on soft skills did not show a link between success and attractiveness (Stinebrickner, 2019). This conclusion is a stark contradiction to previous research and it reveals a novel explanation why attractiveness might be a perk in the job market. Attractive individuals can behave differently as a collective than their less attractive counterparts because of their confidence and self esteem. An analysis of 300 video pitches has revealed that more attractive participants had a higher sense of power and showed a greater nonverbal presence. Both of these factors contributed to more optimal rating of hirability by the participants of the study (Tu, 2020). This means that one can influence his job prospects, regardless of level of attractiveness, by exuding more confidence and working on his self-esteem and soft skills.

While the beauty premium is well established and confirmed, a new research has emerged fully contradicting the stereotype and its existence. A deep look into National Longitudinal Survey of Adolescent Health (Add Health) has shown a direct opposite of beauty premium phenomenon. As discussed by the study, "very unattractive" respondents always earned significantly more than unattractive respondents and sometimes more than average looking respondents. This finding negates the beauty premium and in fact supports a seemingly non existent ugliness premium. The difference in earnings in the study was however explained by individual non physical differences. Health, intelligence, extraversion and lack of neurotic behaviors were the strongest determinants of levels of income (Kanazawa, 2018). Collectively older and a number of recent studies have strongly supported the existence of the beauty premium. The idea that cognitive aspects influence

the beauty premium has been previously rejected. Until recently the benefits of attractiveness were attributed to looks alone. However newer research casts a doubt over the beauty premium and some studies question whether it even exists anymore. While it is possible that with time the beauty premium phenomenon underwent social changes and is not the same as it was before, it is too early to state that beauty premium is less potent and prevalent today. More research has to be done on the topic to conclude just how much beauty premium is about looks and how much the success is determined by cognitive state of mind related to one's looks.

2.2. Most important facial features

We have substantial knowledge that while attractiveness is largely driven by the appearance of health and is strongly influenced by cross-cultural differences, it is finally determined by the environment in which we grow up and our own experiences. This means that while everyone shares common cues in terms of their preference, in the end perception of attractiveness is a subjective matter. In our study we are trying to predict attractiveness scores based on preferences of our participants. What we need to understand is which facial features play the most important roles in determining attractiveness.

While no studies have directly measured differences in subjective ratings of attractiveness of participants on a larger scale, a number of studies have explored aggregate preferences of the participants. Shen and Chau (2006) explored the relationship between the alignment of facial features and the brain responses of the participants using an MRI. By using a software called "Original Face" researchers derive a face that represents the average characteristics of arbitrarily chosen famous Japanese, Korean and Chinese women. Through Original Face the researchers then were able to manipulate the facial features the model face. The software enables to move facial features around and create different views.

In the end the researchers used 432 of the face different variations to present to the participants through manipulation of landmark points on the face. There were in total 29 landmarks available, such as top point of the eyebrows, eye corners, middle point of the mouth, top of the head, bottom of the head and others. The resulting faces differed in the length and width of the face, the location and widths/lengths of eyebrows, eyes, mouth, nose. The goal was to create a broad range of faces with different facial characteristics without creating faces that seem too unrealistic.

The participants were shown the 432 faces in total in 4 rounds. On each showing, their brain activity was scanned using an MRI. They were asked to rate every picture on a scale from 1 to 4, 1 signifying “highly unattractive” and 4 signifying “highly attractive”. The researchers analyzed the participants fMRI results and were able to deduce 6 main components which best determined the attractiveness of the faces shown.

The first and the sixth component were highly related to the nose width and the interocular distance (the distance between the eyes). Second and fourth component were related to the ratios of mideye distance to the interocular distance and nose width. Component 1, 3, 5 and 6 were strongly related to ratio of lip-chin distance and interocular distance. Through component analysis the researchers singled out that the most important facial features are interocular distance, nose width and lip-chin distance.

A study with a more straightforward approach omitting brain signals factor was conducted by Baudouin and Tiberghien (2004). Eight males between the age of 21 and 27 volunteered to participate in an experiment where each male had to rate 62 photographs of real women. The participants were shown one face of a woman on the left side of the screen and a face of another woman on the right side of the screen. The task was to select that face which the participant found to be more attractive. In the end every face in the dataset was compared to every other face in the dataset. This means that each participant was shown 1891 comparisons in total.

The most attractive faces were singled out as those that were chosen in the comparisons the most times, and the least attractive ones were the faces that were chosen the least. The resulting dataset consisted of faces that were both chosen all 61 times and those that were chosen 0 times.

All of the faces were later ascribed 53 face landmarks based on which a variety of features were derived. The features were directly connected to averageness, symmetry, distances between features and area of facial features. The strongest determinant of attractiveness in the study was averageness. There was a strong link between attractiveness and symmetry, but based on the results it was deducted a face had a lower score not because it was asymmetrical, but because asymmetry was a deviation from the average characteristic. Nonetheless, the most attractive faces did not achieve highest score merely because they were close to the average, but because some of their features stood out from the rest. This was true for individuals with more prominent and highly set cheekbones, thicker upper lip and mouth and a small nose (Baudouin, 2004). This conclusion about

symmetry was reinforced by a study conducted by Schmid, Marx and Samal, which focused on attractiveness prediction based on symmetry, golden ratios of facial features and neoclassical canons of beauty to predict attractiveness. Both golden ratios and averageness were stronger predictors of attractiveness than symmetry. While Schmid agrees with Tiberghien that smaller nose widths increase female's attractiveness, Schmid additionally concludes that men have a preference for slender female faces and smaller chins.

A full list of neoclassical canons of beauty includes (Schmid, 2008):

1. Forehead height = nose length = lower face height,
2. Nose length = ear length,
3. Interocular distance = nose width,
4. Interocular distance = right or left eye fissure width,
5. Mouth width = 1.5 x nose width,
6. Face width = 4 x nose width.

In turn a full list of golden ratios includes (Schmid, 2008):

- Ear length to interocular distance
- Lips-chin distance to nose width
- Lip height to nose-mouth distance
- Ear length to nose width
- Interocular distance to eye fissure width
- Length of face to width of face
- Mideye distance to interocular distance
- Interocular distance to lip height
- Nose-chin distance to lip-chin distance
- Mideye distance to nose width
- Nose width to eye fissure width
- Nose width to nose-mouth distance
- Mouth width to interocular distance

- Nose width to lip height
- Mouth width to nose width
- Lips-chin distance to interocular distance
- Eye fissure width to nose-mouth distance

The neoclassical canons of beauty and the golden ratios along with what we have learned from research will help us in defining our own variables that will grasp the measurements of each face in our study.

While the studies so far discussed attempted to understand factors of attractiveness with pre-determined variables, Ibanez-Berganza, Amico and Loreto (2019) allowed the participants of their study to efficiently select their own modification of a face. Each of the participants was asked to align landmarks on a face according to their preferences. The resulting faces represent the ideal settings of facial features according to each of the participants. The resulting faces lead to conclusion that the most important determinants of attractiveness were horizontal and vertical coordinates of facial features. This further confirms the importance of golden ratios. The second conclusion drawn was that while overall facial features created by the participants were close to the average of all faces, singular traits deviated from the average depending on the participant. For example while the measurements and locations of eyebrows, nose and mouth were average, the eye size was set to higher than average. This conclusion is with an agreement with Baudouin and Tiberghien.

In conclusion, in order for our dataset to fully address the preferences of the participants in our study, it is important that we include faces that are close to the average of a population, but also faces that have extreme characteristics and deviate from the norm. This means that faces in our experiment need to exhibit the entire spectrum of attractiveness ratings to provide valid average benchmarks. Secondly golden ratios will serve as cue points of what ratios and facial feature comparisons need to be included to best represent the faces in our dataset. Lastly we need to combine both horizontal and vertical features in our variables.

Our main research hypothesis assumes that based on the characteristics of faces we can accurately predict their subjective attractiveness. In addition, we claim that the attractiveness

assessment is subjective. In comparison to the studies discussed above, we propose an innovative approach by using machine learning algorithms of facial landmark detection to obtain the characteristics. This does expose our study to error due to imperfections of facial detection algorithms. The facial characteristics in the research so far all relied on manually measured facial characteristics, or like in the study by Shen and Chau (2006), a software was used that provided perfect data without any errors. Additionally we show in our study an innovative approach in using a multitude of datasets incorporating a variety of ethnicities instead of relying on just one data source. Lastly, this is the first study that attempts to create multiple predictive attractiveness models based on individual preferences of participants in our study. We verify our main hypothesis by assessing the applied models based on four metrics of classification evaluation (accuracy, precision, recall and f1).

3. Dataset Description

The first step of the study was to collect the pictures of faces for user ratings. Since the study is based on the use of facial landmarks, a consistency between pictures was necessary. Only pictures with neutral facial expression were chosen. This means that exaggerated smiles and other expressions of emotions were avoided. Secondly only pictures with faces looking straight at the camera were picked. If the faces were tilted, the calculations based on the landmarks would be off. Thirdly, since mouth width and eye fissure were factors in the golden ratios, it was necessary to obtain only pictures with closed mouths and open eyes, with some degree of mouth opening allowed. Lastly, it was necessary to include racial diversity in the dataset to avoid underrepresentation of a particular race and to provide our participants full opportunity to define their preferences.

Since no previous studies have explored subjective attractiveness models with multiple participants, the total number of faces had to be large enough to capture the participant's preferences, but not to exceed a number that would overwhelm the raters. In total 862 pictures have been collected from different sources. Due to errors in landmark detection (such as misalignment of mouth, eyebrow, outline of face points) performed by the Dlib library, 181 faces have been removed. The final dataset consisted of 681 pictures.

The pictures were obtained from four sources. The first source was Pinterest. Pinterest is

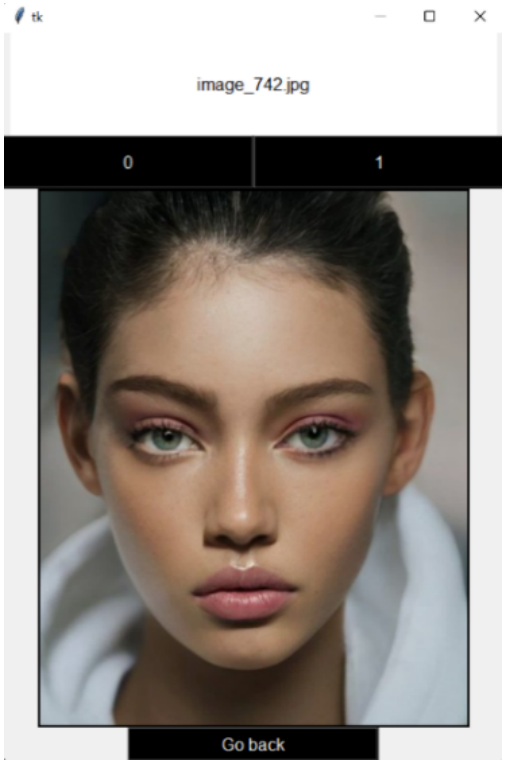
a image based website, where based on the interest of the user, the profile adjusts the images shown. This means the website learns what the user is interested in and shows relevant content. Through typing “neutral face image”, “female face”, “face girl”, “female headshot” and other variations. A total of 389 pictures was collected from Pinterest. The second source was the University of Chicago face database, a high resolution standardized photographs of male and female faces ranging from 17 to 65 years old (Correll, 2021). Out of all available pictures in the database 102 were chosen. The third source was a French project called The Origins of Beauty. This project specializes in finding women from various ethnic groups and aims to show facial differences between races, cultures and ethnicities and provides high quality headshots of women and provided 211 pictures. (Ivanova, 2022). The fourth source was a Chinese dataset providing high quality of famous and non famous women from China (Xie, 2015). Only 20 pictures were chosen from this dataset.

The final dataset consists of 681 images of faces and for each image there are 53 different variables which best represented the face. The full list of the variables and the description of facial landmarks detection is provided in the methodological section of the paper. Since symmetry was not a strong determinant of attractiveness, only two symmetry related variables were included (comparison of eye areas and lengths). 23 variables were related to areas of facial features and their relationships. This means we derived an area of one facial feature or multiple and divided by another. 25 variables were related to distances. Although horizontal and vertical distances have been reported as most important, we also included variables that combine both horizontal and vertical facial characteristics. Finally 3 variables combined distances and areas.

Our dataset consists of female faces only. When it comes to obtaining high quality pictures for such a study, female faces are easier to find. Additionally, if we were to investigate male attractiveness, beards should be included. Unfortunately current facial detection algorithms do not work well with beards. Since our dataset consisted strictly of images of female faces, 5 heterosexual male participants (participants numbered 1 through 5), aged from 26 to 38 years old took part in the rating process. The attractiveness of each face was assessed by each of the participants in two ways: binary with values 0 or 1 (appealing to the participant or not appealing to the participant) and quintary with values from 1 to 5 (1 meaning the lowest attractiveness and 5 indicating the highest attractiveness according to the participant).

The rating was performed using Tkinter library for Python. Tkinter enables users to create simple and interactive user interface. An exemplary view of the rating interface is shown in Figure 1.

Figure 1. An exemplary view of the rating software used



The participants had to press the appropriate button according to their preference, and after the button was pressed the picture changed to the next face. At the end their submissions were saved to the dataset. In total there were 10 datasets meant for classification purposes.

The distributions of attractiveness classes based on the binary participant ratings can be seen in Table 1 and the quintary ratings can be seen in Table 2. The participants' id number can be observed on the leftmost columns.

Table 1. The distributions of classes based on binary ratings

Participant / rating value	0 (non-attractive)	1 (attractive)
1	34.20%	65.80%
2	39.70%	60.30%
3	38.90%	61.10%
4	56.90%	43.10%
5	68.00%	32.00%

Table 2. The distributions of classes based on quintary ratings

Participant/rating value	1 (least attractive)	2	3	4	5 (most attractive)
1	5.50%	12.00%	34.50%	13.00%	35.00%
2	3.40%	25.80%	26.70%	31.80%	12.30%
3	9.00%	23.30%	22.70%	19.30%	25.80%
4	28.20%	22.70%	18.00%	18.00%	13.10%
5	5.00%	21.80%	32.00%	24.00%	17.30%

Table 1 and Table 2 show clear differences in class distributions. In Table 1 respondent 1 found 34.2% of the dataset images unattractive and 65.8% attractive, while respondent 5 found 68.0% unattractive and only 32.0% percent attractive. Table 2 presents cases where some of the classes are significant minorities. While respondent 4 found 28.2% of the dataset most unattractive, the rest of the respondents all rated less than 10% of the dataset as unattractive. On the other hand, respondent 1 gave rating of 5 to 35.0% of faces, while respondents 2, 4 and 5 attributed 5 to at most 17.3%. In addition we checked Pearson correlations for a quintary scale (Figure 2) and Cramer's V statistics for a binary scale (Figure 3) as measures of the strength of relationship (i.e. similarity) between the assessments of different participants.

Figure 2. Pearson correlations between the quintary assessments of attractiveness of different subjects (participants)

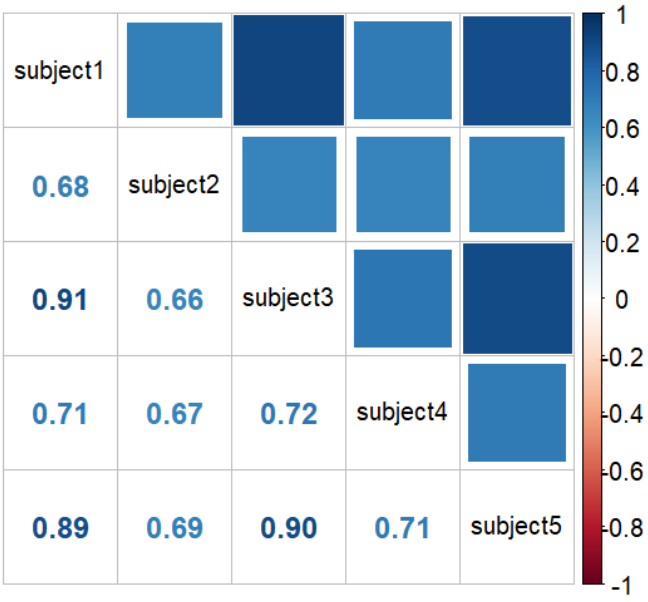
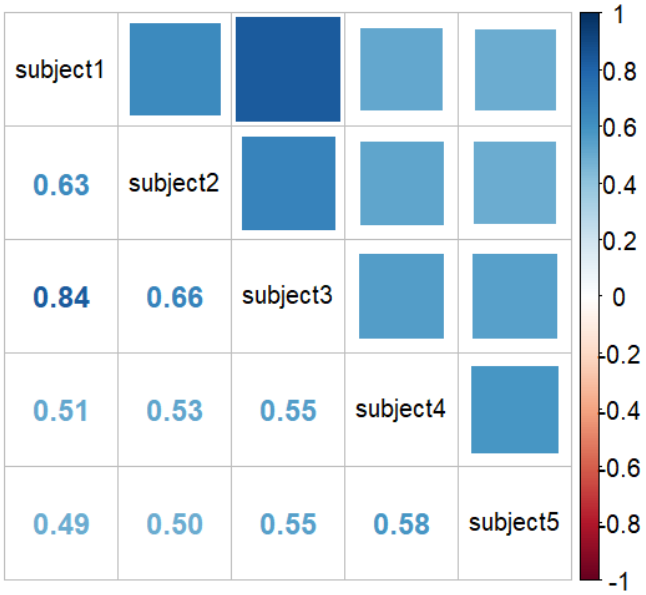


Figure 3. Cramer's V statistics between the binary assessments of attractiveness of different subjects (participants)



Based on Figure 2 one can clearly see that although the assessment of attractiveness on a quintary scale is quite similar for participants 1, 3 and 5 (correlation around 0.9), the remaining correlations are much lower. The differences in assessments (their subjectiveness) are even stronger visible for a binary scale (Figure 3), where attractiveness seems to be similarly assessed only by participant 1 and 3 while it is not so close for all the other pairs of participants. This visible variety in class distributions and differences in assessments show a strong support for subjectiveness and reveal individuality in attractiveness preferences which initially supports our supplementary research hypothesis.

4. Methods

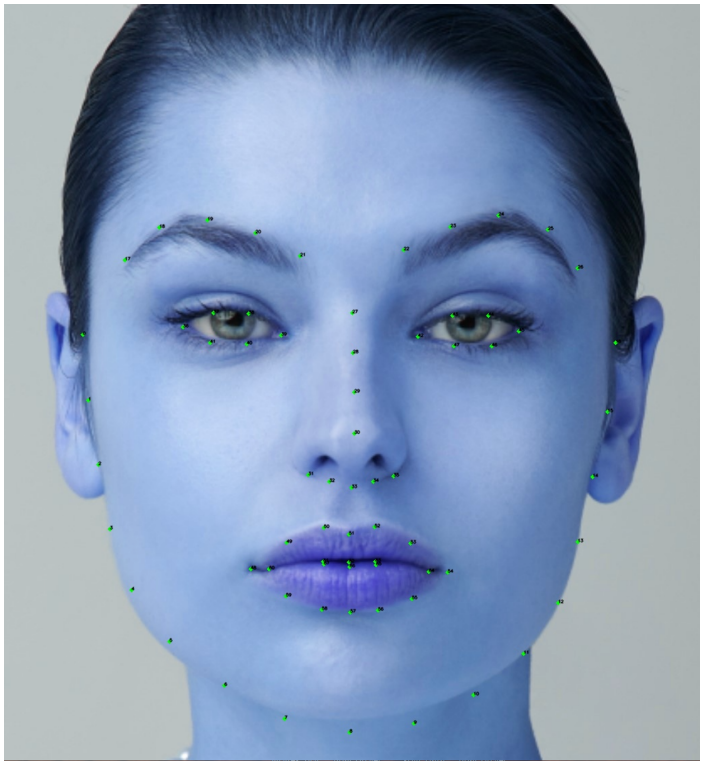
To derive the landmark coordinates from the faces in our dataset we needed a library that provides accurate landmark detection, but also provides a generous amount of landmarks to operate on. One of the most popular and most reliable libraries used for facial landmark detection is the Dlib library. Dlib facial landmark detector provides 68 different points of the face and has proven to be more accurate than other landmark detectors such as STASM (Active Shape Model) (Pool, 2018). The landmarks can be labeled on the face and later accessed to derive the necessary ratios that are needed.

They refer to the following elements of the face:

- inner face area – area surrounding facial features (eyebrows, eyes, nose, lips),
- features area – combined area of eyebrows, eyes, nose and lips,
- inner triangle area – area surrounding lips and eyes,
- left outer face and right outer face areas – areas of cheeks,
- vision area - area surrounding eyes and eyebrows.

For example, if we want to obtain length of an eyebrow, we calculate it by Euclidean distance between landmark at the start of the eyebrow, and the landmark at the end of the eyebrow. An exemplary face with all 68 landmarks is presented on Figure 4.

Figure 4. An exemplary view of a face with 68 landmarks detected by dlib library



Based on the 68 landmarks we created the following 53 variables that describe the face and relations between different features. The features show relationships between the most important elements of the face (eyes, nose, lips, eyebrows) as well as the contours of the face.

symmetry metrics

s1 = eye_area_symmetry – eye area divided by left eye area

s2 = eye_length_symmetry – right eye length divided by left eye length

area ratios

a1 = eyes_to_lips – combined areas of eye divided by lip area

a2 = eyes_to_nose – combined eye area divided by nose area

a3 = eyes_to_face – combined eye area divided by area of the face

a4 = eyes_to_top_face – combined eye area divided by upper portion of face

a5 = lips_to_nose – lip area divided by nose area

a6 = lips_to_face – lip area divided by face area

a7 = lips_to_lower_face – lip area divided by lower portion of area

a8 = nose_to_face – nose area divided by face area

a9 = eyes_lips_nose_to_face – combined areas of eyes, lips and nose divided by face area

a10 = eyes_lips_nose_to_inner_face – combined areas of eyes, lips and nose divided by inner *face area*

a11 = features_to_triangle – features area divided by the inner triangle area

a12 = features_to_face – features area divided by face area

a13 = features_to_outer – features area divided by cheek areas

a14= inner_to_outer – cheek area divided by area between the features

a15 = inner_to_face – inner triangle area divided by face area

a16 = outer_to_face – cheek areas divided by face area

a17 = top_face_to_face – upmost portion of face divided by face ae

a18 = upper_face_to_face – upper portion of face divided by face area

a19 = lower_face_to_face – lower portion of face divided by face area

a20 = bottom_face_to_face – lowest portion of face divided by face area

a21 = upper_to_bottom – upper portion divided by lowest portion

a22 = vision_to_face – vision area divided by face area

a23 = nose_lips_eyes_to_features_area – area of features to the span of area they take up

distances

d1 = face_length_to_width_top – length of face to width of face at top

d2 = face_length_to_width_bottom – length of face to width of face at bottom

d3 = face_top_to_bottom_width – bottom face width to width of face at top

d4 = eye_distance_to_face_width – distance between eyes to width of face

d5 = eyebrows_to_face_width – total length of eyebrows to width of face

d6 = mouth_to_eye_distance – length of mouth to distance between eyes

d7 = mouth_to_eye_spread – mouth length to span of eyes

d8 = mouth_to_nose_width – length of mouth to width of the nose

d9 = nose_length_to_face_length – nose length to face length

d10 = bottom_distance_to_face – chin to mouth over face length

d11 = nose_to_mouth_to_face – nose to mouth distance over face length

d12 = eye_distance_to_nose_width – nose width to eye distance

d13 = nose_length_to_nose_to_chin_distance – nose length to nose to chin distance

d14 = nose_width_to_nose_mouth_distance – nose width to nose mouth distance

d15 = mouth_length_to_width – length of mouth to mouth width

d16 = features_length_to_face_width – total length of eyebrows, eyes, mouth over width of face

d17 = features_length_to_face_length – total length of eyebrows, eyes, mouth over face length

d18 = nose_width_to_eye_fissure – nose_width over eye_fissure

d19 = eye_fissure_to_nose_mouth_distance – eye fissure over mouth to nose distance

d20 = chin_mouth_distance_to_nose_width – mouth to chin distance over nose width

d21 = lips_chin_to_nose_width – lips chin distance to nose width

d22 = lips_chin_to_eye_distance – lips chin distance to eye distance

d23 = nose_width_to_face_width – 4 times nose width over face width

d24 = mouth_width_to_nose_width – mouth width over one and a half nose width

d25 = interocular_distance_to_mouth_fissure – distance between eyes to mouth fissure

distances and areas

da1 = eyebrows_to_eyes – eyebrow lengths over eye areas

da2 = vision_to_face_width – vision area to face width

da3 = features_length_to_features_area – total length of features to area of features

For example variable d12 describes the ratio of eye distance to nose width. This means we calculated the distance between the eye corners (Euclidean distance between two landmarks) and divided by nose width (also Euclidean distance between two landmarks).

The data was randomly divided into train (80%) and test (20%) sets. Models are trained on the training set and their performance is evaluated on the test set. Due to imbalance in the data a sampling transformer augments the underrepresented entries in the dataset and populates the dataset to equally represent all classes. Inefficient representation of a given group in a dataset can lead a model to incorrectly classify representants of that minority class. As Table 1 and Table 2 have shown, there are cases of strong minorities. The oversampling technique used in this study is SMOTE, or Synthetic Minority Over-sampling Technique (Chawla, 2002).

Since there are 5 participants in our study and each one of them is completing two rating processes, this means that we will have 10 total sets of attractiveness preferences. Therefore 10 models in total will be fitted and evaluated.

For the benchmark model for classification we chose logistic regression. Logistic regression models have been used in attractiveness prediction studies with proven significance (Garza, 2016). In addition, we apply tree based models – one based on a bagging technique, and another that is based on a boosting technique. Random forest classifier is a supervised learning algorithm that uses a bagging technique (Breiman, 2001). The bagging approach relies on creating multiple trees that run in parallel without interaction. Each tree is trained on a different randomly selected sample of our dataset. Since each tree is based on a random sample this mechanism helps to avoid overfitting. The boosting algorithm we use is the Extreme Gradient Boosting algorithm (Chen, 2016). While xgboost is also a tree based model unlike in the random forest, xgboost iteratively attempts to improve model on each run. On each iteration the algorithm provides higher weights to those observations that were not fitted well in the previous iteration. Additionally xgboost is capable of performing L1 and L2 regularization. L1 regularization is a type of regularization that discourages the model from using too many features. This is possible by adding a loss function proportional to the sum of absolute values of the coefficients in our model. Therefore the penalty forces the coefficients to be small and in turn the model uses fewer features. The L2 regularization on the other hand also encourages usage of small weights of coefficients, but the loss function operates based on sum of squared values of the coefficients. This forces the coefficients to be small as well, but does not encourage zero weight coefficients. An excellent description and comparison between the L1 and L2 regularization can be found in Ng (2004).

Both, the xgboost and the random forest classifier have hyper-parameters to tune. One of the methods helpful with hyperparameter optimization for modelling is Grid Search (Bergstra, 2012). Grid search allows for testing all at once multiple values for multiple parameters, and based on the results achieved returns the optimal settings of hyperparameters. We apply 10-fold cross validation to find the optimal values of hyperparameters. Full set of hyperparameter settings for a particular model of the participant can be found in Appendix A.

We aim to compare the performance of our models by calculating six different classification models evaluation metrics. The first metric is accuracy. Accuracy tells us what percentage of predictions are accurate. It provides an overall performance of the model, but other metrics have to be included, since accuracy alone can be misleading in case of non-balanced data. In a case where a majority class is predicted 100% correctly and a minority class is predicted 100%

incorrectly, the accuracy score might not reflect the inability of the model to correctly classify the minority class. For that reason we will include main classification metrics precision score, recall and an f1 score metric. Precision defines how good the model is at predicting positive values, or in other words, out of those predictions attributed to a particular class, how many of them actually belong to that class. Recall on the other describes how many predictions are correct out of all the predictions that should have been made. Lastly an f1 score works as a balance between recall and precision. Depending on the balance of the classes we look at these four measures differently. We will also include two additional metrics, specificity and balanced accuracy for extra precision in our analysis. Specificity determines a model's ability to predict whether an observation does not belong to a particular category. Balanced accuracy is similar to overall accuracy, but it considers the imbalances of the datasets.

The evaluation of the best models will be completed with an analysis of the four previously discussed metrics (accuracy, precision, recall, f1 score), but also confusion matrices. A confusion matrix visualizes the model's predictions on a map. The visualization of predictions allows us to see where our model wrongly predicted our classes. The advantage of confusion matrix analysis over metrics is that it shows directly where the predictions are.

Additionally we will analyze which facial features are the most important through feature importance feature. This means we will know which features influence the performance our model the most. The importance of a given variable is calculated as the mean and the standard deviation of accumulation of impurity decrease within each tree.

5. Empirical Analysis

Considering a large number of 53 variables used and facial features repeatedly being used in the calculations of the variables, it was necessary to check the correlations between the independent variables (see Table 3). A high correlation between the variables can lead to overfitting the model. For our threshold for correlation between variables we chose 0.9 coefficient. We found pairs of variables with correlations surpassing our threshold and removed the ones that conflicted with most variables (for example if we had three variables a, b, c, and variable a would have a correlation of 0.9 and above with b and c, but b and c would have a lower correlation between them, we would remove variable a) . After removing the most correlated variables, there were 47 remaining in total.

The removed variables were A4, A7, A9, D2, D21, D22. The correlated variables can be observed in Table 3.

Table 3. Most correlated predictors

Variable 1	Variable 2	correlation
A4	A3	0.98
A7	A6	0.93
D2	D1	0.94
D20	D21	1.00
D16	D5	0.95
D13	D9	0.92

The removed variables were A4, A7, D2, D20, D16, D13.

After splitting the dataset into testing and training test and after applying SMOTE was applied to our training set, three different models (elastic net, xgboost and random forest) were fitted and tested on all 10 datasets (5 binary and 5 quintary). A precision score was derived from every model to represent it's goodness of fit. An average was taken for each group to determine which type of model performs the best for attractiveness prediction task. The best performing model for both, the binary and quintary datasets was xgboost. The results for the binary models can be observed in Table 4, and the results for quintary models can be observed in Table 5.

Table 4. Performance measures for the binary models on the test sample

	Elastic net						Random forest						xgboost					
	accuracy	balanced accuracy	recall	specificity	precision	F1	accuracy	balanced accuracy	recall	specificity	precision	F1	accuracy	balanced accuracy	recall	specificity	precision	F1
1	0.82	0.83	0.83	0.83	0.80	0.81	0.80	0.81	0.80	0.81	0.77	0.78	0.84	0.83	0.84	0.83	0.81	0.80
2	0.82	0.81	0.82	0.81	0.82	0.82	0.84	0.81	0.80	0.81	0.83	0.83	0.84	0.85	0.85	0.84	0.84	0.80
3	0.81	0.80	0.80	0.79	0.79	0.81	0.86	0.86	0.90	0.86	0.85	0.86	0.87	0.87	0.87	0.86	0.86	0.90
4	0.79	0.80	0.80	0.79	0.79	0.79	0.82	0.85	0.80	0.85	0.83	0.82	0.85	0.86	0.86	0.85	0.85	0.90
5	0.77	0.76	0.77	0.75	0.75	0.76	0.81	0.77	0.80	0.77	0.79	0.78	0.82	0.81	0.77	0.81	0.80	0.80
avg	0.81	0.80	0.80	0.79	0.79	0.80	0.83	0.83	0.80	0.82	0.81	0.81	0.84	0.84	0.84	0.84	0.83	0.80

Normally we would analyze precision scores, recall, F1 and accuracy separately to compare our models since these metrics react differently depending on sizes of datasets and the balance of the classes, but since xgboost outperforms random forest and elastic net models on every evaluation metric we do not have to do that and xgboost is a clear winner in terms of performance. The average values for the four evaluation metrics oscillate around 0.84. There are no outliers in terms of performance. This means that our models can effectively capture the attractiveness preferences of our participants. the evaluation metrics for each of the participants in xgboost are close to each other, with the biggest variance in results in participant 5, where recall was 0.05 points lower than accuracy. This means that model of participant 5 had a higher tendency to falsely identify faces as a particular class.

Table 5. Performance measures for the quintary models on the test sample

	Elastic net						Random forest						xgboost					
	accuracy	balanced accuracy	recall	specificity	precision	F1	accuracy	balanced accuracy	recall	specificity	precision	F1	accuracy	balanced accuracy	recall	specificity	precision	F1
1	0.54	0.87	0.57	0.53	0.50	0.49	0.61	0.90	0.58	0.59	0.55	0.56	0.61	0.89	0.55	0.54	0.54	0.54
2	0.44	0.84	0.42	0.47	0.33	0.33	0.42	0.84	0.50	0.51	0.42	0.42	0.47	0.85	0.49	0.46	0.48	0.48
3	0.45	0.86	0.49	0.47	0.43	0.43	0.50	0.87	0.52	0.52	0.49	0.49	0.50	0.87	0.50	0.47	0.48	0.49
4	0.45	0.85	0.43	0.41	0.42	0.41	0.41	0.86	0.38	0.41	0.36	0.36	0.47	0.86	0.45	0.40	0.44	0.44
5	0.48	0.87	0.49	0.50	0.46	0.47	0.47	0.86	0.45	0.41	0.47	0.45	0.55	0.87	0.54	0.49	0.59	0.56
avg	0.47	0.86	0.48	0.48	0.43	0.43	0.48	0.87	0.49	0.49	0.46	0.46	0.52	0.87	0.51	0.47	0.51	0.50

With our quintary models xgboost also outperforms random forest and elastic net in precision, recall, f1 and accuracy. Nonetheless, the performance is worse than in the binary case. While the average accuracy for binary models and the average accuracy was 0.52. Additionally the variance in performance of the models was larger. The best performing model scored 0.61 (participant 1) in accuracy, while the worst performing model scored 0.47 (participant 4). This means that preferences of participant 1 were clearer for the xgboost model than the preferences of participant 4 were. It is possible that participant 4 was less consistent in terms of characteristics than participant 1 was. For example, if participant 4 included in the class 5 (most attractive) female faces with both wide and narrowly set eyes, and did the same for class 1 (least attractive), such set of preferences could confuse the model. The metrics of precision, recall, f1 and accuracy were similarly close to each other like in the binary models.

Lower performance on the quintary datasets is understandable. The number of faces needed to fully capture someone's preferences has not been explored and the number is an unknown. Considering we included faces from different datasets and also included different ethnicities (Afro-American, Asian, Caucasian, Eurasian and others) it is probable more data is needed to create more effective models. Considering participants expressed difficulty in choosing appropriate values for displayed face in quintary ratings, it is also probable our models would benefit from a more elaborate

The hyperparameters tested through Gridsearch were number of estimators, colsample_bytree, learning rate, reg_lambda and reg_alpha. The full spectrum of parameters tested can be found in appendix A. The resulting best parameters output by gridsearch for each of the models can be observed in Table 6 and Table 7.

Table 6. Set of best hyperparameters for binary xgboost models

Participant/parameter	Learning rate	Colsample_bytree	N_estimators	Reg_lambda	Reg_alpha
1	0.3	0.8	200	1	0.2
2	0.3	0.8	200	1	0.2
3	0.3	0.5	200	1	0
4	0.3	0.5	300	1	0
5	0.3	0.5	300	1	0

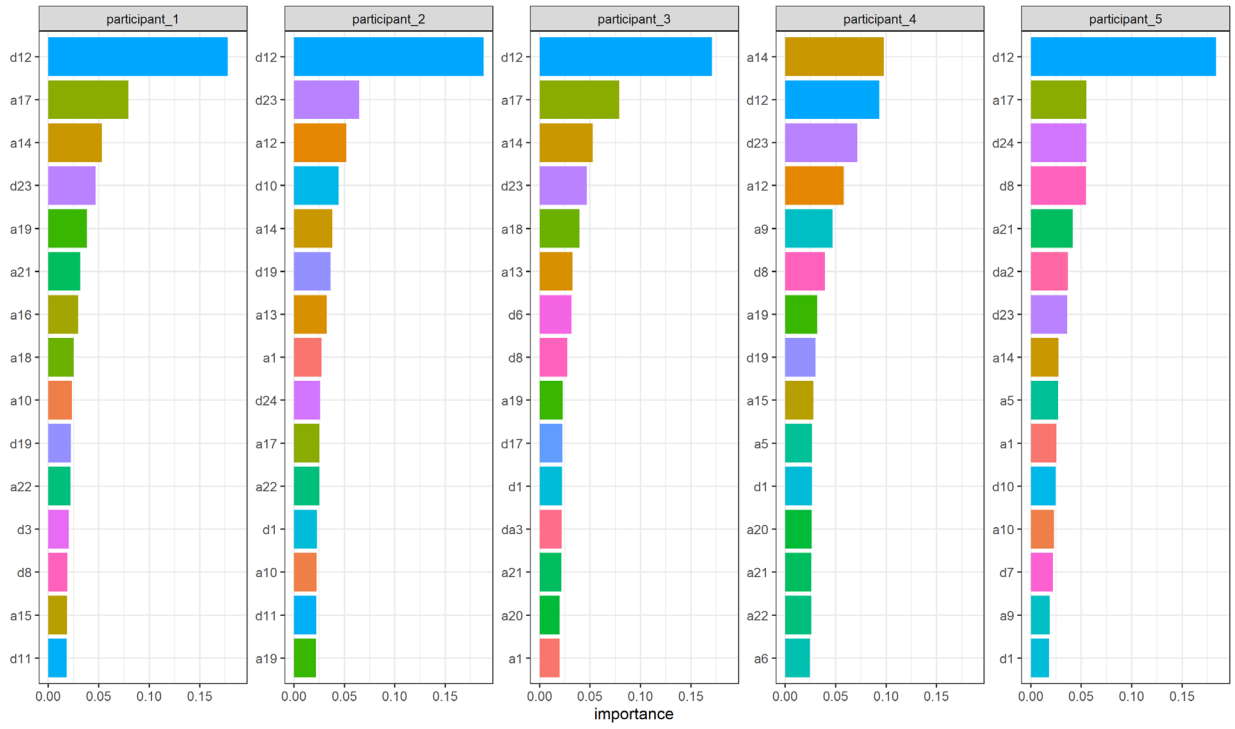
Table 7. Set of best hyperparameters for quintary xgboost models

Participant/parameter	Learning rate	Colsample_bytree	N_estimators	Reg_lambda	Reg_alpha
1	0.3	1	200	0.8	0.1
2	0.3	1	200	0.8	0.5
3	0.3	0.5	300	0.8	0.5
4	0.3	0.5	300	0.8	0.1
5	0.05	0.5	200	0.8	0.5

The main difference between the quintary and binary xgboost models can be observed in the subsample parameter. The default value of the parameter (subsample equal to 1) was preferred for the binary classification, while lower values were appropriate for the quintary classification. This

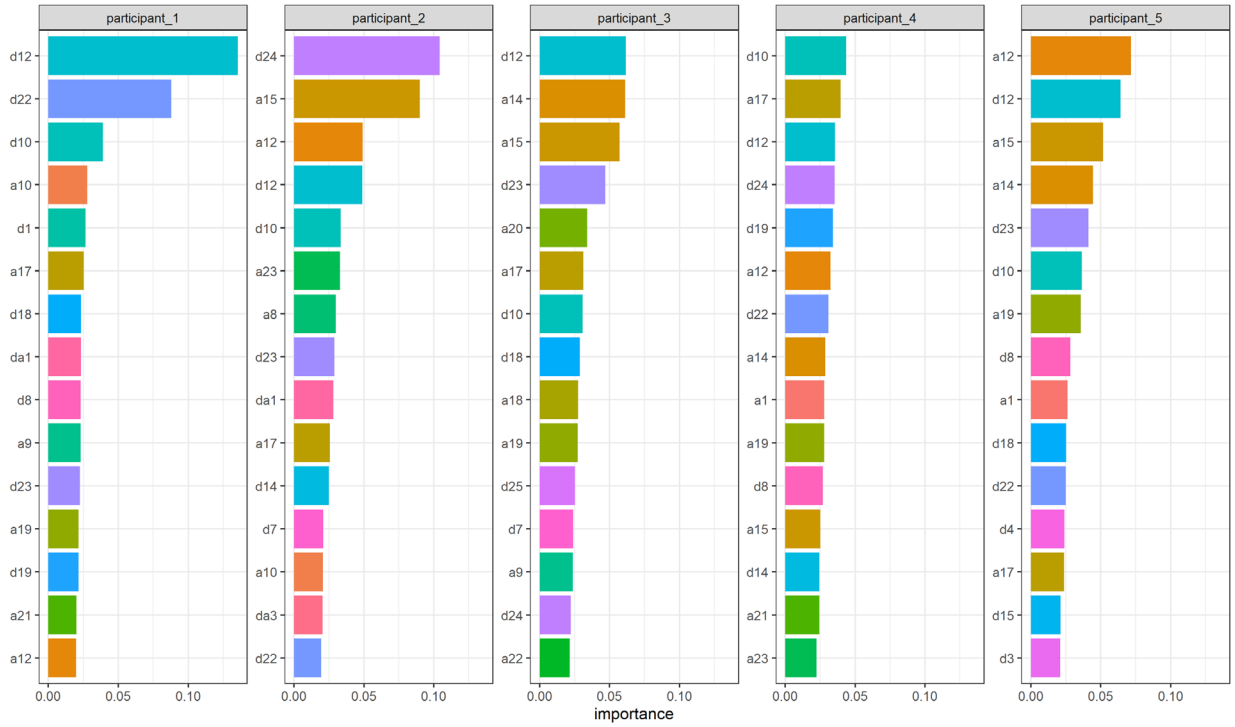
makes sense since lower values of subsample are better suited when model is susceptible to overfitting. With more classes and higher class imbalance, fitting models on quintary data poses overfitting risks.

Figure 5. Feature importance based on binary models (top 15 features for each participant)



To understand the differences between the assessment of attractiveness by different participants, we look at the feature importance metrics for all 5 binary models in Figure 5. Across all participants the feature that consistently achieves high score of relevance is the variable d12, which is distance between the eyes divided over nose width. Besides participant 4, for all other four participants the d12 variable was unquestionably the most important predictor of attractiveness. Other recurring significant variables were a17 (total length of eyebrows, eyes, mouth over face length), d23 (4 times nose width over face width). Interestingly there are variables important for one participant, but not important for other participants. Those are, for example, variable d10 for participant 2 and a14 for participant 1, 3 and 4. While all 5 participants have common important features, there are also clear differences that reveal subjectiveness in their ratings.

Figure 6. Feature importance based on quintary models (top 15 features for each participant)



If we look at the feature importances for quintary models in Figure 6 we can notice that once again the highest relevance was collectively achieved by the variable d12. The variable d12 was either of highest importance or nearly the most important. Variable a15 was found highly important for participants 2, 3 and 5. Variable d10 was important for all participants, ranking in the top 8 positions for all participants.

Nonetheless despite a common factor of the nose width across the models it can be argued that differences in feature importances, class imbalances and performances of the models strongly support the idea that attractiveness is a subjective matter. There is no clear pattern in the feature importance graphs and while in 1, 2, 3 and 5 the disparity between the most important features and the rest of the features is evident, model 4 shows a much more balanced importance across all features.

6. Conclusions

We claimed that based on characteristics of a face we can accurately predict the subjective attractiveness of a face. To verify our hypothesis we asked 5 participants to rate 681faces on two scales – binary and on a scale from 1 to 5. We created xgboost classification models based on randomly selected training data and tested the models on the testing data. This process was done for each of the participants and 10 models were created in total (5 for binary classification and 5 for quintary classification). We evaluated our models with four metrics (accuracy, precision, recall and f1) that together describe how well a model is performing. While our hypothesis is confirmed in binary case and combining user preferences with facial features derived with the use of landmarks and processed by machine learning algorithms holds up, the same cannot be said for the quintary case. It is possible that the sample of 681 faces is not enough to fully grasp preferences of our raters. A possible solution would be to increase the sample, or only include one ethnicity in the sample the variety of facial features.

This is one of the first studies, where multiple models are simultaneously built to predict subjective attractiveness. There are many extensions of this research possible. Firstly, with such abundance of variables and overlapping features (such as eyes, nose, eyebrows, mouth) it is possible that the amount of variables could be reduced by applying PCA, or principal component analysis. This means that multiple features could belong to one vector and be represented by that vector alone. A PCA analysis could reveal in more depth which features and their ratios are most relevant, and which are redundant and should be ignored. Additionally considering 68 variables returned by dlib and how many combinations can be used, it is possible that 47 variables are insufficient to represent the face.

In our study we used the dlib landmark detector tool to detect faces. At the moment state-of-the-art facial landmark detectors are still prone to error and cannot consistently detect facial landmarks with pinpoint accuracy. A study with manually drawn landmarks would ensure that the data used in the study is of highest quality and reduce the measurements errors. Additionally, if one feature is miscalculated by a landmark detector, that means multiple variables in the study are incorrect and the whole data entry can be misinterpreted by our models. A follow up study with manually performed landmark marking could improve the legitimacy of the model.

Considering that participants have expressed difficulty in choosing appropriate rating values

for a number of faces in the quintary rating process, it is possible to extend the study by creating a more in depth rating process to ensure the datasets are of high quality and faithfully represent the preferences of our participants.

References

Anýžová, P., & , P. (2018). Beauty still matters: The role of attractiveness in labour market outcomes. *International Sociology*, *33*(3), 269–291.

Baudouin, J. Y. & Tiberghien, G. Symmetry, averageness, and feature size in the facial attractiveness of women. Acta Psychologica 117, 313–332 (2004).

Breiman, L. (2001) Random Forests. University of California, Berkeley, Statistics Department.

Bergstra, J., Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 13 (2012) 281-305

Bronstad, P. M., Russel, R. (2007). Beauty is in the 'we' of the beholder: greater agreement on facial attractiveness among close relations. Perception. 2007; 36: 1674-1681

Brooks, R., Scott, I. M., Maklakov, A. A., Kasumovic, M. M., Clark, A. P., & Penton-Voak, I. S. (2011). National income inequality predicts women's preferences for masculinised faces better than health does. Proceedings of the Royal Society of London B. 278, 810–812.

Chawla, V. N., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002) SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321-357

Chen, T., Guestrin, C. (2016). XGBoost: A scalable Tree Boosting System.

DeBruine, L., M. (2004). Resemblance to self increases the appeal of child faces to both men and women. Evol. Hum. Behav. 2004; 25: 142-154

Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is what is good. Journal of Personality and Social Psychology, 24, 285-290.

Duorui Xie, Lingyu Liang, Lianwen Jin∗, Jie Xu, SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception, Mengru Li School of Electronic and Information Engineering South China University of Technology, Guangzhou 510

Fletcher, J.M., 2009, Beauty vs. brains: early labor market outcomes of high school graduates, Economics Letters 105, 321-325.

French, M.T., 2002, Physical appearance and earnings: further evidence, Applied Economics

34, 569-572.

Garza, R., Heredia, R. R., & Cieslicka, A. B. (2016). Male and Female Perception of Physical Attractiveness: An Eye Movement Study. *Evolutionary Psychology*, *14*(1)

Gupta, N. D., Etcoff, N. L., Jaeger, M. M. (2016) Beauty in Mind: The Effects of Physical Attractiveness on Psychological Well-Being and Distress. Journal of Happiness Studies 17, 1313-1325 (2016)

Hamermesh, D.S. and J.E. Biddle, 1994, Beauty and the labor market, American Economic Review 84, 1174-1194.

Hamermesh, D.S., Meng, X., and J. Zhang, 2002, Dress for success - does primping pay?, Labour Economics 9, 361-373

Ibanez-Berganza, M., Amico, A., Loreto, V. (2019) Subjectivity and complexity of facial attractiveness. Sci Rep 2019; 9: 8364

Ivanota, N. (2022). The project The Ethnic Origins of Beauty. https://lesoriginesdelabeaute.com/fr/accueil

Jokela M. (2009), "Physical attractiveness and reproductive success in humans: Evidence from the late 20 century United States", Evolution and Human Behavior, vol. 30(5):342-350, https://doi.org/10.1016/j.evolhumbehav.2009.03.006.

Kanazawa, S. and Still, M. C. (2018). Is There Really a Beauty Premium or an Ugliness Penalty on Earnings? Journal of Business and Psychology, 33(2), 249–262, https://doi.org/10.1007/s10869-017-9489-6

Little, A. C., Bur, D. M., Perrett, D. I. (2006). What is good is beautiful: face preference reflects desired personality. Pers. Individ. Dif. 2006; 41: 1101-1118

Ma, D., Correll, J., Wittenbrink, B. (2021). Chicago Face Database. The University of Chicago. https://www.chicagofaces.org/

Manning, J. T., Anderton, R. & Washington, S. M. J. Hum. Evol 31, 41-47 (1996).

Møller, A. P. (1990). Parasites and sexual selection: Current status of the Hamilton and Zuk hypothesis. Journal of Evolutionary Biology, 3, 319–328.

Nault K.A., Pitesa M. and Thau S (2020), "The attractiveness advantage at work: A cross-disciplinary integrative review", Academy of Management Annals, vol. 14(2), pp. 1103-1139, https://doi.org/10.5465/annals.2018.0134

Nithya, S. (2015). Dove's Choose Beautiful Campaign: A Marketing Misadventure?. Amity Research Centers.

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance.

Computer Science Department, Stanford University, Stanford CA 94305, USA.

Pfeifer, Ch. (2011). Physical Attractiveness, Employment, and Earnings. Leuphana University Lunenburg. Discussion Paper No. 5664.

Pflüger L.S. et al. (2012), "Cues to fertility: perceived attractiveness and facial shape predict reproductive success", Evolution and Human Behavior, Volume 33, Issue 6, pp. 708-714, https://doi.org/10.1016/j.evolhumbehav.2012.05.005.

Pool, W. (2018). Comparison of various landmark detecting techniques in the context of forensic facial recognition. University of Twente. EEMCS

Ralph Stinebrickner, Todd Stinebrickner, Paul Sullivan; Beauty, Job Tasks, and Wages: A New Conclusion about Employer Taste-Based Discrimination. *The Review of Economics and Statistics* 2019; 101 (4): 602–615.

Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, S. M., Hightower, A., & McKay, R. (2001). Do facial averageness and symmetry signal health? Evolution and Human Behaviour, 22(1), 31–46.

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration and facial attractiveness. Psychological Science, 7, 105–110

Scholz JK, Sicinski K. (2015). Facial attractiveness and lifetime earnings: evidence from a cohort study. Rev Econ Stat. 2015 Mar;97(1):14-28. doi: 10.1162/REST_a_00435. Epub 2015 Mar 2.

Shen, Hui., Chau, K. P. Desmond., Su, Jianpo. (2016). Brain responses to facial attractiveness induced by facial proportions: evidence from an fMRI study. Scientific Reports, volume 6, Article number 35905.

Schmid, K., Marx, D. & Samal, A. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. Pattern Recognition 41, 2710–2717 (2008).

Singh, D. & Zambarano, R. J. Hum. Biol. 69, 545-556 (1997)

Stephen, Ian., & Wei, T. (2015) Healthy body, healthy face? Evolutionary approaches to attractiveness perception

Stephen, I. D., Coetzee, V., Perrett, D. I. (2011). Carotenoid and melanin pigment coloration affect perceived human health. Evolution and Human Behavior, 32(3), 216–227.

Swami, V. & Tovée, M. J. (2005). Female physical attractiveness in Britain and Malaysia: A cross-cultural study. Body Image, 2, 115–128.

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. Human Nature, 4, 237–269.

Tovée, M. J., Reinhardt, S., Emery, J. L., & Cornelissen, P. L. (1998). Optimum body-mass index and maximum sexual attractiveness. Lancet, 352(9127), 548–548.

Tu, M.H., Gilbert, E.K., & Bono, J.E. Is beauty more than skin deep? Attractiveness, power, and nonverbal presence in evaluations of hirability. Personnel Psychology. 2021; 1–28.

Yu, Douglas & Shepard, Glenn. (1999). reply: The mystery of female beauty. Nature. 399. 10.1038/20348.

Xie, D., Lian, L., Jin, L., Xu, J. (2015). SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception. Mengru Li School of Electronics and Information Engineering South China University of Technology, Guangzhou.

Appendix A

XGBoost hyperparameters grid:

1. number of estimators: [100, 200, 300]
2. learning rate: [0.05, 1, 2, 3]
3. subsample: [0.5, 0.8, 1]
4. lambda: [0, 0.1, 0.5]
5. alpha: [0, 0.5, 0.8, 1]
6. colsample by tree: [0.5, 0.8, 1]

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl