16. FEBRUAR 2018

PREDICTING PERSONALITY TRAITS FROM SOCIAL MEDIA FOOTPRINTS - AN ENHANCED MODEL

Erstprüfer: Prof. Dr. Burkhardt Funk
Zweitprüfer: Prof. Dr. Henrik von Wehrden

NIKLAS MAXIMILIAN MRUTZEK
M.SC. MANAGEMENT & DATA SCIENCE
Martrikelnummer: 301 9124

TABLE OF CONTENS

Abbreviations

EN – Elastic Net

LASSO – Least Absolute Shrinkage and Selection Operator

PCC – Pearson Correlation Coefficient

RBF – Radial Basis Function

SVM – Support Vector Machine

Tables

Figures

INTRODUCTION

Evaluating another person’s personality is an essential part of human life. How an individual reacts to a certain trigger, let it be a statement, strongly depends on his personality. Therefore, knowledge about the personality of a conversational counterpart is crucial to predict how he or she will react to a question or an answer. Personality is commonly understood as “*patterns of thought, emotion, and behavior that are relatively consistent over time and across situations*” (Funder 2012). If personality is as aforementioned defined as stable “over time and across situations”, then it has to be differentiated from the character, which might change as an actor plays a role. A large proportion of an individual’s outer behavior can be explained by the inner personality. The outer behavior as a result of the personality determines various socio-demographic attributes, like job satisfaction (Furnham et al. 2002), the success of romantic relationships (Noftle, Shaver 2006), job performance (Barrik, Mount 1991) or high income, conservative political attitudes, early life adjustment to challenges, and social relationships (Soldz, Vaillant 1999). Humans can infer another person’s personality pretty precise. A first impression like a short video in many cases is enough to asses a personality (Carney et al. 2007). However, personality assessment is not limited to the social-cognitive domain of human brains – machine learning models attempt to predict personalities as well, or even better than humans. The internet provides a vast amount of data regarding personal information about its users – to so-called digital footprint. Especially social networks offer personal data in a very condensed form, the social-media footprint. Social media networks, which are online platforms, where people create a profile of themselves and communicate with other users or artificial persons like newspaper, offer a wide range of personal data to the broad community, as well as the network and its developers. In the year 2014 49.7 % of the German internet participated in social media networks (Statistisches Bundesamt 3/16/2015) with an upward trend. Furthermore, social media networks, like Facebook, provide the possibility to “like” something, which means at first: the user starts to follow a certain page and therefore receives updates and messages from the page and secondly: that the user publicly declares that he or she likes the page, visible to other users. However, it has been shown that the profile of a social network user indeed reflects the individual user and his personality and not an “idealized” version of themselves (Back et al. 2010). Hence, these profiles seem to be unbiased, or at least as biased as the personality tests themselves.

On the other side are the Facebook pages. A page in this case can be related to anything that a user started, let it be a political attitude, an artificial person, a company or a special kind of food. Any page can be created, and every user can give it a “Like”. Facebook, as the biggest social media network as of today (Statista 2017) offers the possibility to collect data about a user’s Facebook likes, if the user agrees to the request. Due to the generic nature of Facebook likes and the relevance of personality assessment as a crucial part of social living, this paper focuses onto machine personality prediction based on Facebook likes. However, listening to music from a certain group in a web browser or reading a certain online newspaper can be easily translated into the Facebook like analogy and vice versa, which means that findings from this study are unlikely limited to the domain of Facebook likes.

PERSONALITY ASSESMENT

Various researchers gave different definitions of what constitutes an individual’s personality. The concept of personality is a merely abstract one, defining underlying traits of an individual’s character. While a character could be consciously adjusted to the actual situation, like an actor plays a role, is the personality of an individual independent of the situation, environment or any kind external influence. Hence, personality describes a merely generic kind of psychological trait set. Because of its independence regarding external influences is a desirable trait of a definition of personality a certain amount of stability over time, cross-cultural stability and – what many models lack – a descriptive model, a taxonomy. The Big 5 taxonomy evolved out of this indifference in the researching community, as it does not represent a certain theoretical argumentation, neither does it replace previous or other perspectives; it is on one hand derived from natural language data and on the other hand, “[…] the Big Five taxonomy serves an integrative function because it can represent the various and diverse systems of personality description in a common framework […]” (John et al. 2008, p.116). The Big Five model has been verified by various studies and it is sensible to claim that it is considered the international standard model regarding personality as it has been used in more than 3000 studies in the last 20 years (John et al. 2008, p.116), a result of its intercultural stability (John, Srivastava) and the attributability to genetics instead of environment (Bouchard, McGue 2003). McCrae et al. found that “heritability, limited parental influence, structural invariance across cultures and species, and temporal stability all point to the notion that personality traits are more expressions of human biology than products of life experiences” (McCrae et al. 2000, p. 177). The Big 5 taxonomy itself, namely Openness, Conscientiousness, Extraversion, Agreeablesness and Neuroticism, as defined by John et al. (2008) can be found in *Table 1 – The Big Five Taxonomy.*

As aforementioned, did the Big Five Model evolve from the idea that individuals express their personality by the way they phrase and their wording. This idea was first formulated by Allport and Odbert (1936) where the authors defined different lexica for the different dimensions of personality, the so called “nomenclature”. However, expressions of personality are neither limited to the sphere of linguistics, as it has been shown that personality can be predicted from spaces people inhabit (Gosling et al. 2002), nor is personality assessment limited to the capabilities of the human brain. Keeping the nomenclature in mind, the first approaches of machine based personality assessment make use of textual data and their linguistic features. Fruyt et al. (2004) used a modern version of the nomenclature to predict the Big Five inventory. So did Fast and Funder (2008), who predicted personality profiles based on samples of written text, furthermore is the individual personality predictable from linguistic features derived from weblogs (Oberlander, Nowson 2006). This indicates that personality assessment is a merely unique human ability.

Table 1 – The Big Five Taxonomy

	Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Verbal labels	Originality Open-Mindedness	Contstraint Controle of impulse	Enthusiasm Energy	Altruism Affection	Negative Emotionality Nervousness
Conceptual definition	Describe the breadth, depth, originality, and complexity of an individual's mental of experiential life	Describes socially prescibed impulse control that facilitates task- and goal-directed behaviour, such as thinking before acting, delaying gratification, following norms and rules, and planning, organizing and proritizing taks	Implies an energetic approch toward to the social and material world and includes ttraits such as sociability, activity, assertiveness and positive emotionality	Contrats a prosocial and communal orientation toward others with antagonism and includes traits such as altruism, tender-mindedness, trust and modesty	Contrasts emotional stability and even-temperedness with neagtive emotionality, such as felling anxious, nervous, sad and tense
Behavioral examples	Take the time to learn something simply for the joy of learning; Watch documentaries or educational TV; Come up with novel setups for my living space; Look for stimulating activities that break up my routine	Arrive early or on time for appointments; Study hard in order to get the highest grade in class; Double-check a term paper for typing and spelling errors; Let dirty dishes stack up for more than one day (R)	Approach strangers at a party and introduce myself; Take the lead in organizing a project; Keep quit when I disagree with others (R)	Emphasize the good qualities of other people when I talk about them; Lend things to people I know (e.g. class notes, books, milk); Console a friend who is upset	Accept the good and the bad in my life without complaining or bragging (r); Get upset when somebody is angry with me; Take it easy and relax (R)
Examples of external criteria predicted	**High pole:** Years of education completed; better performance on creativity tests; success in artistic jobs; create distinctive-looking work and home environments **Low pole:** Conservative attitudes and political party preference	**High pole:** Higher academic grade-point averages; better job performance; adherence to their treatment regimens; longer lives **Low pole:** Smoking, substance abuse, and poor diet and exercise habits; attentiondeficit/ hyperacticity disorder (ADHD)	**High pole:** Social status in groups and leadership positions; selection as jury foreperson; positive emotion expression; number of friends and sex partners **Low pole:** Poorer relationships with parents; rejection by peers	**High pole:** Better performance in work groups **Low pole:** Risk for cardiovascular disease, juvenile deliquency, interpersonal problems	**High pole:** Poorer coping and reactions to illness; experience of burnout and job changes **Low pole:** Feeling committed to organizations; greater relationship satisfaction

RELATED WORK

First attempts of algorithmic personality predictions using data from social media profiles, were also based on language and derived linguistic features (Mairesse et al. 2007). Personality traits do not just reveal themselves in the wording of an individual, but also in the number of friends in a social network. Bai et al. (2012) used social media profiles from the most popular Chinese platform RenRen to predict personality. They enriched their linguistic features with socio-demographic, geographic data and the user's RenRen network density to improve prediction accuracy. Bachrach et al. (2012) tested various possible features based on Facebook data, for example network density, number of uploaded photos, number of group memberships etc. Also browsing logs (Goel et al. 2012) were shown to be predictive of personality, as well as personal online blogs (Marcus et al. 2006) and collections of music that users listen to (Rentfrow, Gosling 2003). Furthermore, personality can be predicted from Facebook friendship network density and number of friends, or from Twitter profiles (Quercia et al. 2011). Even the location within the Facebook friendship network was shown to be predictive of the user's sexual orientation (Jernigan, Mistree 2009). Major research regarding predicting personality from Facebook Likes comes from Stanford University researchers Kosinski et al. (2013) and Youyou et al. (2015). Kosinski et al. (2013) applied singular value decomposition and linear regressions onto a dataset of *N=55.814* Facebook users and their respective Facebook likes to predict socio-demographic details and four of the Big Five personality traits. Using n=170 Facebook likes per user on average did the method peak at a Pearson Correlation Coefficient of Agreeableness 0.3, Extraversion 0.4, Consciousness 0.29, and Openness 0.43. Youyou et al. (2015) represent state of the art research. The team used Least Absolute Shrinkage and Selection Operator (LASSO) to select the relevant Facebook Likes for a linear regression. Using an estimate of n=227 Facebook Likes on average for a user did the method peak at a Big Five average accuracy of PCC~0.56 and showed a log-linear relationship between the number of Facebook Likes and prediction accuracy. As aforementioned, this model can be seen as the state of the art solution to predict personality based just on Facebook Likes, therefore this method was used as a baseline performance indicator for this paper.

DATA

The dataset used in this study, which is the same as used by Youyou et al. (2015), was obtained from the myPersonality project. The project was hosted at Stanford University and offered psychometric tests and feedback on their scores, while collecting the respective user's Facebook Likes. The sample used for this study consists of *N=10000* instances of the five personality dimensions and the user's corresponding Facebook Like structure. The following chapter provides descriptive statistics and information about correlations and non-independence in the data.

Descriptive Statistics

TARGET VARIABLES

Table 2 - Descriptive statistics of the Target Variable (N=10000)

	Openness	*Consciousness*	*Extraversion*	*Agreeableness*	*Neuroticism*
Mean	4.02	3.63	3.61	3.60	2.71
Std	0.63	0.70	0.82	0.68	0.81
Min	1.25	1.00	1.00	1.00	1.00
25%	3.65	3.20	3.00	3.25	2.10
50%	4.10	3.70	3.75	3.70	2.75
75%	4.50	4.15	4.25	4.08	3.25
Max	5.00	5.00	5.00	5.00	5.00

The Big 5 dimensions are measured in an interval {1,5} and they tend to correlate. A data visualization can be found in *Figure 4: Target variables boxplot* and *Figure 8 - Target variable (paired co-) density* plots in the appendix. Note that the distribution of the dimensions Consciousness, Extraversion and Agreeableness are very close to each other.

EXPLANATORY VARIABLES

The explanatory variables in this case are the like structure of the Facebook User. The dataset consists of 93.871 possible Facebook Likes and 10000 instances. Note that users with less than 20 Facebook Likes have been excluded. *Table 3 - Descriptive statistics of Like structures* shows that the Facebook Like structures are very imbalanced from both perspectives: Number of Likes per user and Number of likes per page. The number of Likes per user varies from 0 to 2807 and the number of Likes per page varies from 0 to 2397, keeping the quantiles in mind, this does indicate a very imbalanced dataset. Therefore, a desirable trait of the model would be a certain level of independence from the amount of Facebook Likes available. Using the Facebook Likes as features (m ϵ M) results in a very high dimensional problem setup as M >> N. Furthermore, the data is extremely sparse with just ~0.08% of available Facebook Likes, or in other words: the user does not like a random page in 99.92% of the cases. Taking all the arguments into consideration, the data is imbalanced regarding amount of likes per user, as well as for pages, it is extremely sparse and very high dimensional.

Table 3 - Descriptive statistics of Like structures

	Likes per User	Likes per Page
Count	10000	93871
Mean	72.26	7.70
Std.	102.37	32.95
Min	20.00	0.00
25%	29.00	0.00
50%	44.00	1.00
75%	75.00	5.00
Max	2807	2397

Correlation and independence

Independence is an important standard assumption in machine learning theory, therefore it is crucial to understand why the features are not independent in this dataset. Two random variables X and Y are said to be uncorrelated when their correlation coefficient is zero:

$$\rho(X,Y) = 0$$

Equation 1

since

$$\rho(X,Y) = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$$

Equation 2

being uncorrelated equals saying the covariance is zero. As

$$Cov[X,Y] = E[XY] - E[X]E[Y]$$

Equation 3

having zero covariance is the same as

$$E[XY] = E[X]E[Y]$$

Equation 4

If the expectation of the product factors equals the product of the expectations, then two variables are uncorrelated. If $\rho(X,Y) \neq 0$ then two variables are (at least slightly) correlated. Another case is the statistical independence. Two variables are said to be independent if their joint probability distribution is the product of their marginal probability distributions.

That means for all x ϵ X and y ϵ Y:

$$p_{X,Y}(x,y) = p_X(x)\, p_Y(y) \qquad \text{Equation 5}$$

This means that the conditional distribution is the same as the marginal distribution

$$p_{X|Y}(y|x) = p_Y(y) \qquad \text{Equation 6}$$

If X and Y are not independent, then they are dependent by definition. The link between correlation and independence is conditional: If X and Y are independent, then they are also uncorrelated, however if X and Y are uncorrelated, then they can still be dependent. This can be proven by showing that:

$$\begin{aligned} E[XY] = \quad & \iint x, y\, p_{X,Y}(x,y)\, dxdy \\ & \iint x, y\, p_X(x)\, p_Y(y)\, dxdy \qquad \text{Equation 7} \\ & \int x\, p_X(x) \left(\int y\, p_Y(y)\, dy \right) dx \\ & \left(\int x\, p_X(x)\, dx \right) \left(\int y\, p_Y(y)\, dy \right) \\ & = E[X]E[Y] \end{aligned}$$

Therefore, it is sufficient to proof that two variables are correlated in order to show that they are not independent. The most common way to show that two random variables are correlated is the Pearson correlation coefficient (PCC). The PCC is common sense to proof two variables are either correlated or uncorrelated and has the null-hypothesis of $\mathrm{H}_0\colon \rho = 0$ and an alternative hypothesis of $\mathrm{H}_A\colon \rho \neq 0$ in the case of a two-tailed significance test. A pairwise combination without repetition of possible Facebook Likes occurring at least 100 times in the dataset has been conducted. The PCC has been calculated on every pair of this combination, showing that on a significance level of alpha=5% the null-hypothesis had to be rejected for 34.66 % of the pairs. Furthermore, nearly all possible Facebook Likes have at least one correlated counterpart (namely 99.91%). Therefore, it is sensible to say that Facebook Likes in general are correlated and that means that they are not independent.

METHOD

The previous chapter showed that the data is extremely sparse, highly imbalanced, high-dimensional and that the features are not independent. The multicollinearity can be handled using non-linear regression algorithms. The two most common families of techniques that are based on fitting the parameters of complex, nonlinear functions are nonlinear support vector machines and neural networks (Provost, Fawcett 2013). As visualized in *Figure 8 - Target variable (paired co-) density plots* (see the Appendix) the target variables nearly follow a normal distribution, hence is it sensible to make use of support vector machines with a Gaussian (radial basis function) kernel to predict the personality traits. Feature selection is used to reduce the dimensionality. Furthermore, support vector machines benefit from feature selection in terms of performance, as well as computational costs.

Feature selection

A common approach when working with high dimensional data is to perform a feature selection. Feature selection is mainly done for the sake of interpretability and to improve prediction accuracy. Especially in regression tasks models often reveal a low bias but a high variance. A common approach to lower the amount of variance in prediction accuracy is to set some coefficients to zero – in other words: the selection of the most important features only. In many cases does the model's bias increase, but the increase is much smaller than the decrease in variance; therefore, overall prediction accuracy increases. In machine learning literature, this is often referred to as the "bias-variance tradeoff". In addition to that, does the model indeed get more interpretable: a model that makes use of many different features is intuitively much harder to understand than a model that just makes use of the most important features. It is sensible to phrase this as a "bias-interpretability tradeoff" in which a model may gains a small of amount of bias but wins a big amount of interpretability (Hastie et al. 2009, p.57).

However, the most important argument to make use of feature selection in this case is that SVMs strongly benefit from feature selection in terms of performance (Guyon et al. 2002, p.15), and computational costs.

Feature selection techniques can be split into two different methods: “hard-thresholding” and “soft- thresholding”. Hard-thresholding is thereby commonly known as subset selection. Hastie et al. (2009, p.61) state:

> The process of subset selection, which is basically a process of finding a subset of the features and ignoring the rest, is interpretable and it can lower the prediction error in many cases. However, it is a discrete process that exhibits variance. Small changes in the threshold can result in big changes of the subset. Shrinkage models, as the Ridge or the LASSO, have a continuous nature and therefore they tend to exhibit less variance than subset selection. One of the most modern feature selection algorithms is the Elastic Net, which combines two desirable traits of the Ridge and LASSO Regression.

The next chapters will introduce the Elastic Net feature selection.

RIDGE REGRESSION

The Ridge regression model is continuous, because it does not explicitly discard features from the model, instead it punishes every coefficient by its magnitude in the Euclidian space. It makes use of soft-thresholding by minimizing the penalized sum of squares with its coefficients. The Ridge regression is defined as follow:

$$\hat{\beta}^{\,ridge} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 ,$$

Equation 8 - Ridge Regression

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$

, where $x \in X$ is the feature set, β_j a coefficient in the regression model, y_i the true value and t the hyperparameter. The smaller *t* is*,* the less features will be selected by the model. Hastie et al. (2009, p.63) state that “a wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficient […], this problem is alleviated”. The *squared* sum of coefficients is at maximum *t*, being the reason for calling this a *L*$_2$ penalty term. The *L*$_2$ constraint keeps the solution of the Ridge regression linear in the y_i, resulting in a closed-form solution and relatively low computational costs. The disadvantage of Ridge regression is that it cannot (at least in practice) set coefficients of features exactly to zero due to the nature of the *L₂* penalty term which imposes a penalty onto coefficients but never sets them exactly zero. Ridge regression imposes a proportional shrinkage onto every coefficient, therefore Ridge regression is not a “feature selector” on its own.

LASSO REGRESSION

The Least Absolute Shrinkage and Selection Operator (LASSO) regression is much like the Ridge regression, but with an important difference in its constraint. It is defined as follows:

$$\hat{\beta}^{\,lasso} = argmin_{\beta} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 ,$$

subject to $\sum_{j=1}^{p} |\beta_j \ | \leq t$

Equation 9 - LASSO Regression

Note that the minimization problem is the same as in *Equation 8* -of the Ridge regression; just the penalty changed from the sum of squared coefficients ≤ t (Ridge) to sum of absolute values ≤ t (LASSO), that means the *L₂* penalty has been replaced by an *L₁* penalty term. The *L₁* penalty term makes the LASSO regression a non-linear minimization problem, which is a quadratic programming problem. However, the *L₁* penalty term makes the LASSO regression more interpretable: imposing a penalty of $t = \sum_{j=1}^{p} |\beta_j \ |$ results in nothing different than an ordinary least squares regression and $t = (\sum_{j=1}^{p} |\beta_j \ |)/2$ in an ordinary least squares regression where all coefficients are shrunk by on average circa 50%.

LASSO regression shrinks the absolute values towards zero, Ridge instead shrinks the sum of squared coefficients. The result is that LASSO provides a sparse solution, as shown in *The relation between the LASSO and Ridge* becomes clearer with their generalized form:

$$\tilde{\beta} \quad = argmin_{\beta} \left\{ \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|^q \right\}$$

Equation 10 - LASSO/Ridge generalization

, where $\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)$ is *the* estimator from *Equation 8* -and Equation 9 - and $\lambda \sum_{j=1}^{p} \left|\beta_j\right|^q$ the generalized constraint. Using this generalized form makes clear that

LASSO and Ridge regression just differ in *the* choice of q. While *LASSO regression* uses q=1 does *Ridge regression* use q=2. One can say that LASSO efficiently eliminates trivial features, but from a group of highly correlated features LASSO tends to pick the most predictive features and shrink the others to zero.

The relation between the LASSO and Ridge becomes clearer with their generalized form:

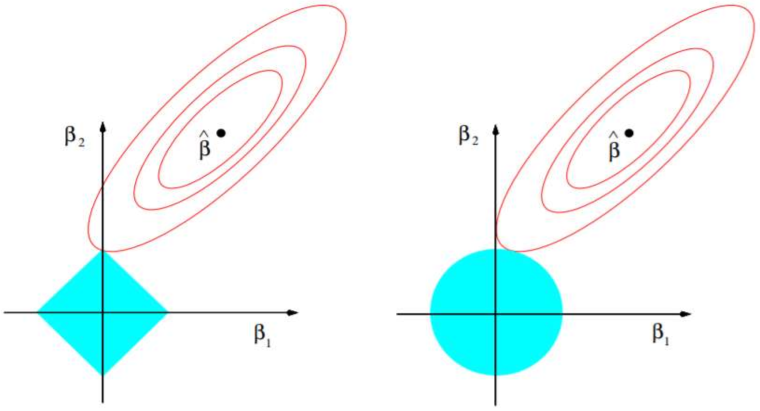
$$\tilde{\beta} \quad = argmin_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|^q\right\}$$

Equation 10 - LASSO/Ridge generalization

, where $\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)$ is the estimator from *Equation 8* -and *Equation 9* - and $\lambda\sum_{j=1}^{p}|\beta_j|^q$ the generalized constraint. Using this generalized form makes clear that LASSO and Ridge regression just differ in the choice of *q*. While LASSO regression uses *q=1* does Ridge regression use *q*=2. One can say that LASSO efficiently eliminates trivial features, but from a group of highly correlated features LASSO tends to pick the most predictive features and shrink the others to zero.

ELASTIC NET

As aforementioned, SVMs benefit from feature selection in terms of performance and computational costs. The LASSO regression performs a feature selection, but it tends to select only the most predictive feature from a group of correlated features and sets all the others zero. The Ridge regression on the other hand does not provide feature selection but it shrinks correlated features towards each other, which is a desirable trait using a SVM with a radial basis function kernel, which is able to make use of interactions between features.



The elastic net (EN) is a compromise between Ridge and the LASSO using the EN penalty of the form:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right),$$

Equation 11 - EN penalty

where first term equals the Ridge penalty and the latter the LASSO penalty. α controls the relative strength of both penalties and is defined $\{0,1\}$, where low α gives more control to the LASSO penalty and a high α more control to the Ridge. Note that as α increases, fewer features will be selected. The estimator of the EN can be setup as *Equation 10* - with the penalty as written in Equation 11 - The EN penalty has the advantage that it does perform feature selection while shrinking correlated features onto each other, without discarding them. One could say that EN shrinks like the Ridge and selects like the LASSO.

Support vector machines and the radial basis function kernel

As shown in *Correlation and independence* the features are correlated and therefore not independent. Hence, non-linear regression models are needed to predict the data correctly. An easy way to perform non-linear regression is the use of a polynomial feature space instead of the original one, but the high-dimensionality of the dataset makes this idea computationally unfeasible. Support vector machines (SVMs) are supervised-learning models that can be used for classification as well as regression tasks and the so called “kernel-trick” allows SVMs to fit complex and non-linear data. Furthermore, the SVM finds, speaking in terms of a classification task, the optimal separating hyperplane and not just *a* separating hyperplane like many other classification algorithms which results in the SVM overperforming compared to less sophisticated algorithms. SVMs maximize the margin of a separating hyperplane. ε-SV regression (Vapnik 2000) seeks to find a function *f(x)* that fits the data with a deviation which is at most ε while being as flat as possible.

The complexity will be increased throughout this chapter; thus, the start is a simple linear function *f* of the form:

$$f(x) = \langle x, \beta \rangle + \beta_0, \qquad \text{Equation 12}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot-product in X. The function shall be flat in order to avoid overfitting, which can be achieved by minimizing the magnitude of β which will be denoted as $\|\beta\|$:

$$\min \frac{1}{2}\|\beta\|^2 \qquad \text{Equation 13}$$

$$\text{subject to } y_i - \langle x, \beta \rangle - \beta_0 \leq \varepsilon \text{ , } y_i - \langle x, \beta \rangle + \beta_0 \leq \varepsilon \text{ ,}$$

where the objective function minimizes the magnitude of β with the constraint that no deviation should be larger than ε. The equation above comes with the assumption that such a function actually exists, which is rarely the case in practice. Therefore, analogous to the "soft-thresholding" method in the chapter Feature selection, it is sensible to introduce the slack variables ξ_i, ξ_i^* for every data point to cope with the optimization problem. From the two slack variables, one relaxes the first constraint and the other one the second constraint from *Equation 13*. The slack variables allow deviations up to in sum C, hence defined as $C \geq \sum_i^N (\xi_i + \xi_i^*)$, where C reflects a tradeoff between allowed deviations and the smoothness of the curve. Using slack variables leads to the so-called ε-insensitive loss function:

$$|\xi|_\varepsilon := \begin{cases} 0 \; if \; |\xi| < \varepsilon \\ |\xi| - \varepsilon \quad else \text{ ,} \end{cases} \qquad \text{Equation 14}$$

which changes the optimization problem in *Equation 13* to the stated one by Vapnik (2000):

$$\min \frac{1}{2}\|\beta\|^2 + C \sum_i^N (\xi_i + \xi_i^*) \qquad \text{Equation 15}$$

$$\text{subject to } \begin{cases} y_i - \langle x, \beta \rangle - \beta_0 \leq \varepsilon + \xi_i \\ y_i - \langle x, \beta \rangle + \beta_0 \leq \varepsilon + \xi_i^* \\ \xi_i \text{ , } \xi_i^* \geq 0 \end{cases}$$

The actual situation is visualized in *Figure* in the appendix. The common approach for finding the global minimum in a constrained optimization problem are Lagrange multipliers.

The introduced Lagrange multipliers $\eta_i \ , \eta_i^*, \alpha_i^*, \alpha_i$ have to satisfy a positivity constraint and the Lagrangian of Equation 15 is:

$$\mathrm{L}_p := \frac{1}{2}\|\beta\|^2 + C\sum_{i}^{N}(\xi_i + \xi_i^*) - C\sum_{i}^{N}(\eta_i \ \xi_i + \eta_i^* \xi_i^*)$$
$$-\sum_{i=1}^{N}\alpha_i \ \left(\varepsilon + \xi_i \ - y_i + \langle\beta, x_i\rangle + \beta_0\right)$$
$$-\sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* - y_i - \langle\beta, x_i\rangle - \beta_0)$$

Equation 16 – The primal Lagrangian of the optimization problem

which is minimized w.r.t. to the primal variables β, β_0 and $\xi_i^{(*)}$. Setting the respective derivates to zero, one gets:

$$\partial_{\beta_0} L = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0$$

Equation 17

$$\partial_{\beta} L = \beta - \sum_{i=1}^{N} x_i(\alpha_i^* - \alpha_i) = 0$$

Equation 18

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

Equation 19

Note that *Equation 18* can be rewritten as $\beta = \sum_{i=1}^{N} x_i(\alpha_i^* - \alpha_i)$, which yields the final values for all $\hat{\beta}$. Substituting Equation 17*, 18* and *19* into *Equation 16* – yields the dual optimization problem, which is maximized to close the duality gap:

$$max\begin{cases} -\frac{1}{2}\sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i + \alpha_i^*) \end{cases}$$

subject to $\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i, \alpha_i^* \ \epsilon \ \{0, C\}$,

Equation 20 – The Dual optimization problem

whose Lagrangian results in a lower bound for the objective function in *Equation 15*.

The derivation of *Equation 20* – eliminated the $\eta_i^{(*)}$ and introduced $\langle x_i, x_j \rangle$, as aforementioned, *Equation 18* be rewritten as:

Equation 21

$$f(x) = \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) \langle x_i, x \rangle + \beta_0,$$

which is the **support vector expansion** (Smola, Schölkopf 2004). It is shown that all β can be explained as a linear combination of the input features from X. As normally the complexity of a dataset can be described by the dimensionality of the input features, does in this case the complexity depend on the number of support vectors. This is one of the main advantages of SVMs compared to other techniques, although not unique to them. Furthermore, the entire algorithm can be described by dot-products of the input features. The latter term of the equation, namely the β_0 can be calculated using the Karush-Kuhn-Tucker conditions for the optimal solution (Kuhn, Tucker 1951).

SVMs fit non-linear dataset basically by preprocessing. In this case, preprocessing means the map the original feature space into an expanded feature space by applying a map function $\Phi: X \rightarrow G$ onto the dataset and then apply the above explained SVM onto *G* to find a linear solution in *G*, which yields a non-linear solution in *X*. One could think of polynomial transformation in the original feature space to perform linear separation within the constructed feature space, which translates to non-linear separation in the original feature space. However, as shown in the chapter *Correlation and independence*, the dependencies in the original feature space are very strong and relate to high polynomial, making the calculation quickly computationally infeasible due to the quadratic growth. The so-called “kernel trick” is a bypass to the computational costs. The expanded feature space can be represented by the inner product of the input features via the kernel trick, which is computationally cheap. As noted before, the SVM algorithm just involves the dot products of x_i. Hence, for an arbitrary mapping function Φ it is sufficient to have knowledge about the kernel $k(x, x') := \langle \Phi(x), \Phi(x') \rangle$, not the mapping function itself.

Therefore, *Equation 20* – can be generalized to the non-linear case:

$$max \begin{cases} -\frac{1}{2}\sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) \\ -\varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i + \alpha_i^*) \end{cases}$$

subject to $\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i, \alpha_i^* \epsilon \{0, C\}$,

Equation 22 – The generalized Dual optimization problem

As well as the support vector expansion in *Equation 21*:

$$f(x) = \sum_{i=1}^{N}(\alpha_i + \alpha_i^*)\, k(x_i, x) + \beta_0$$

Equation 23

The optimization problem moved into another feature space *G* by Φ. Loss is minimized in *G*, not in *F*, resulting in a model complexity that just depends on the number of support vectors. (Smola, Schölkopf 2004) summarize:

> “The input pattern (for which a prediction is to be made) is mapped into feature space by a map Φ. Then dot products are computed […] of the training patterns under the map Φ. This corresponds to evaluating kernel functions *k(xi, x)*. Finally, the dot products are added using the weights $\hat{\beta}_i = \alpha_i - \alpha_i^*$. This, plus the constant term β_0 yields the final prediction output. The process described here is very similar to regression in a neural network, with the difference, that in the SV case the weights in the input layer are a subset of the training patterns.”

A very popular kernel among kernel methods is the radial basis function (RBF) kernel (also called Gaussian kernel). It has the form:

$$K(x, x') = \exp(-\gamma\|x - x'\|^2),$$

Equation 24 - The RBF kernel

where γ is a free parameter that has to be chosen (hyperparameter). The RBF kernel projects the input space into an infinite dimensional feature space, as the RBF kernel can be understood as an infinite sum over polynomial kernels. Generally spoken, the kernel reflects a measure of similarity between two vectors. The similarity in the RBF-kernel is measured using the squared norm distance (squared Euclidian distance) of the two vectors x , x'. If two vectors are close to each other, then $\|x - x'\|^2$ is small, therefore $-\gamma\|x - x'\|^2$ is large (if $\gamma > 0$). This results in a bell-shaped curve, where γ controls the width of the curve. The width of the curve decreases with increased γ, leading to the situation that values farther away from the regression line will be taken into consideration. As SVM complexity is driven by the number of support vectors, a very high gamma generally leads to overfitting, while very small values for γ make the model too constraint (underfitting).

Taking all the aforementioned arguments into consideration, one can see that the *C* in combination with γ are critical for a SVM regression to achieve a good perfomance.

RESULTS

The results of the state of the art LASSO regression (Youyou et al. 2015), as discussed in *Related work,* are visualized in *Figure* in the appendix. Using an N=10.000 and an average number of Facebook Likes of 72.26 the LASSO regression peaks at a ***PCC~0.35***, using 10-fold cross validation, predicting extraversion. The prediction accuracy regarding extraversion is generalizable, confirming the findings of Youyou et al. (2015), keeping the smaller training set of N=8,000 in mind, compared to the N~63,000 in the original study.

The accuracy of elastic net feature selection and SVM regression is shown in and is the result of 10-fold cross validation, where 20% of the data is used as the test set, resulting in a training set of n_{train}=*8000*. The standard deviation between the test scores of the folds is on average ~0.013, which indicates stable model performance. Accuracy is measured using the PCC, which is defined in an interval {-1,1}, where PCC=0 implies no correlation. The PCC as an evaluation metric is chosen to ensure comparability to the aforementioned state of the art model. The correlation of predicted and real value is statistically significant for all Big Five dimensions p<0.0001****. The average number of Facebook Likes in the data (72.26) corresponds to an accuracy of PCC~0.68. Youyou et al. (2015) reported 227 as the average number of Facebook likes per individual with a corresponding accuracy of PCC=0.56, whereas proposed EN feature selection and SVM regression predicts with an accuracy of PCC~0.76, which is achieved with a ten times smaller training set. Furthermore, correlation is much higher compared to an average human personality judge (*PCC = 0.49* (Youyou et al. 2015)), showing that the model outperforms human judges decisively.

All hyperparameters have been tuned using grid search. Along with the findings of (Guyon et al. 2002), SVM regression benefit from feature selection. However, the optimal number of features as an input for the SVM regression in this case is around *M~5580* across the Big Five dimensions, compared to the fine-tuned alpha hyperparameter of pure LASSO regression with *M~453*. The resulting feature selection of the EN for the SVM regression can be interpreted as a very weak feature selection, where every feature that is somewhat related to the target variable is selected. Knowledge of the RBF-kernel reveals that the SVM regression also benefits from features that are not predictive by themselves, but in combinations with others. For an initial hyperparameter search, grid search followed a logarithmic grid of 20 equidistant values, and were fine-tuned farther on a second grid search attempt on a smaller interval. The resulting hyperparameters are equal across all personality dimensions, indicating equal importance and observability from the data with $C = 100\,, \gamma = 0.001, \varepsilon = 0.05.$

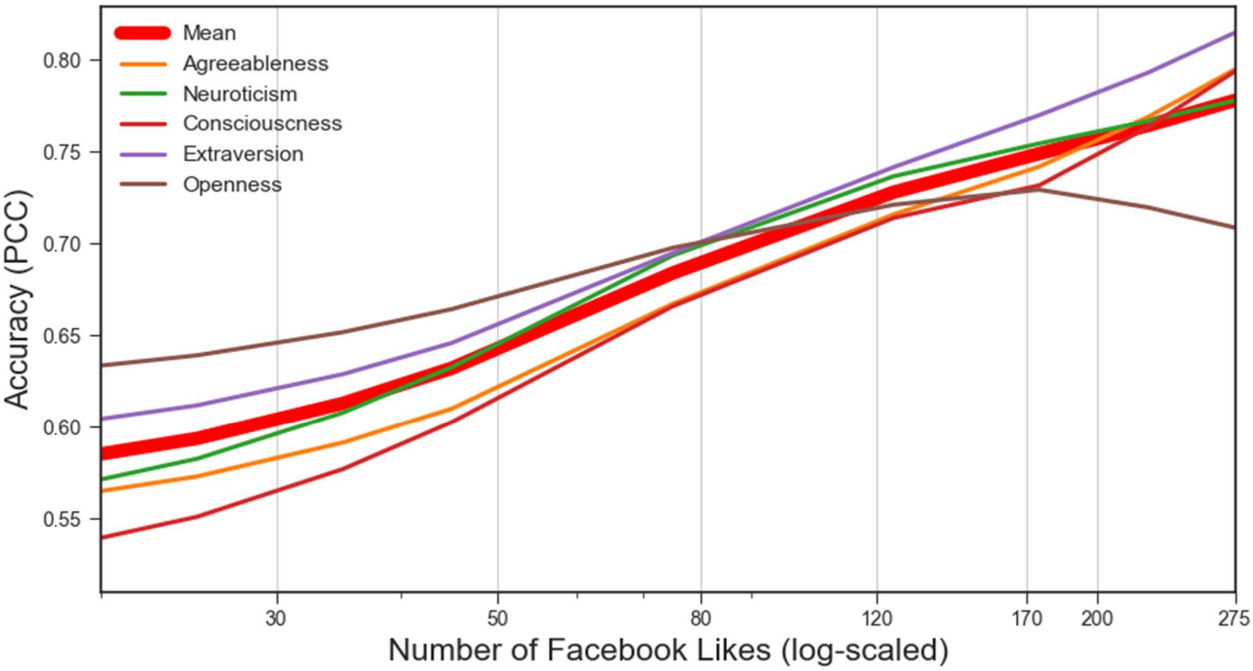


Figure 2 – EN feature selection and SVM regression accuracy, measured as PCC, depending on the log-scaled number of available Facebook Likes per individual. The accuracy curve has been smoothed using a LOWESS approach.

Grid search pattern for the critical hyperparameters is γ and C is visualized in *Figure 3*

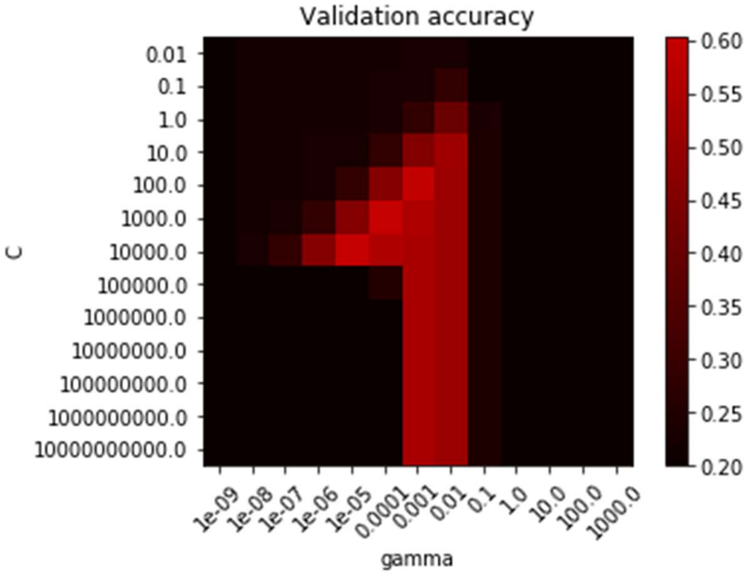


Figure 3 - Grid search heatmap. Prediction accuracy as PCC depending on γ and C. High accuracy shows a (weak) linear relation between the hyperparameters. The data is exemplary for predicting extraversion, but the pattern is generalizable.

The proposed method shows better performance, especially compared to the LASSO regression, in the cases where just a small amount of information is present, namely a small number of Facebook likes from an individual. The method achieves an average accuracy of *PCC=0.57* with just *n=15* Facebook likes, which is comparable with the personality judgement accuracy of the corresponding individual's spouse (*PCC=0.57,* (Youyou et al. 2015)). This can be related to an accuracy stabilizing effect due to the use of a non-linear regression model, which is able to make use of correlated variables instead of dropping them. Furthermore, deviation in prediction accuracy across the Big Five dimensions is reduced, which can be understood as another stabilizing effect.

However, a drawback of the proposed method are the computational costs compared to LASSO regression. The fitting SVMs with an RBF-kernel comes with high computational costs, mainly because of the very complex feature transformation in the kernel. Especially the grid searching procedures with five hyperparameters (L_1/L_2-ratio and α for EN, C, γ and ε for the SVM regression), is very time consuming. Furthermore, fitting time strongly depends on sample size, hence training on a smaller sample size is recommended. Computations for this paper were made with a 46-core machine as cloud computing service. Despite that, the calculation of *Figure 3* took ~4.5 hours. However, the scoring time of a SVM regression is much lower than the one of LASSO regression, as SVMs can store the support vector inside the model. One more practical consideration is to make use of the SciPy package in python, which offers a special kind of dataframe, the scipy.sparse dataframe. Classical dataframes store every “non-like” in the matrix as a zero, which results in massive memory requirements. Furthermore, grid-searching is nearly infeasible on a local machine, as one fitting process of the SVM-regression with a sample size of N=10,000 takes about 30 seconds. Grid-searching ten different values of the three hyperparameters of the SVM-regression within a 10-fold cross-validation setup, results therefore in computational costs 10^4 fitting processes. This results in ~ 83 hours of computation, using one core. Using the 46 cores cloud-computing machine thus results in total fitting time of around two hours. Fitting a sample size of N=40,000 takes approximately three minutes, resulting in total fitting time of twelve hours for the aforementioned cloud-computing instance.

DISCUSSION

The proposed method shows superior performance to human personality assessment, as well as to the state of the art method. Knowledge about just 15 Facebook likes of an individual allow the model to make more precise statements about the personality than the individual’s closest human counterpart, namely the spouse. First, the method shows that Facebook likes are very rich data to predict personality, likely due to their generic nature of just expressing that an individual “likes” something. Second, SVM regression strongly benefits from selecting a broad range of anyhow predictive features (in terms of EN: very small alpha), to transform these into a much larger and complex feature space (in this case make use of an RBF-kernel) and make predictions based on features that might be not predictive by themselves but in combination with others. Third, the personality of a Facebook user with a lot of likes can be extremely accurately calculated with the method peaking at *PCC~0.8*, which is, statistically spoken, overall an unreached precision of external personality assessment based on Facebook Likes.

SVM regression with a RBF-kernel is a black box algorithm, which means, as opposed to white box algorithms, that they do not deliver any reasoning behind the decision they make. Hence, it is not possible to say which Facebook like is predictive for e.g. a highly-extroverted individual. This means that no reasoning can be delivered from the method itself. However, research suggests possible methods of white boxing the RBF-kernel (van Belle, Lisboa 2014). Next to white boxing, another future improvement could break the binary setup of Facebook likes (either an individual liked a page, or he or she did not). The assumption of binarity does not hold in practice, as a “non-like” can be interpreted in two ways: either the user really does not like the page, or he just does not know about it. The method presented in this paper comes with the assumption that the user does not like the page; what obviously might not be true. Approaches from the field of Recommender Systems could be a solution to this problem. While content-based filtering would just bootstrap into the data what one already knows from the original dataset, would collaborative filtering be able to make predictions about “non-likes” if the user really does not like the page or if he would like it if he would know about it.

As useful as these methods are for business, especially marketing – thinking of advertisements fitted to the consumer’s personality – machine-based personality prediction comes with ethical considerations. Users might stop to use social media networks when they understand them as intruders into privacy, as digital assessment comes to maturity. In terms of predictive power what comes from the data, policy-makers have to be aware of data privacy. Some studies suggest that it is possible to manipulate people by having a knowledge about their personality, as it has been shown by Hirsh et al. (2012), or during the last presidential election in the United States (Grassegger, Krogerus 2016).

PUBLICATION BIBLIOGRAPHY

Allport, Gordon W.; Odbert, Henry Sebastian (1936): Trait-names. A psycho-lexical study. Princeton, N.J, Albany, N.Y: Psychological Review Company (Psychological review publications, 211 = 47,1).

Bachrach, Yoram; Kosinski, Michal; Graepel, Thore; Kohli, Pushmeet; Stillwell, David (2012): Personality and patterns of Facebook usage. In Noshir Contractor, Brian Uzzi, Michael Macy, Wolfgang Nejdl (Eds.): Proceedings of the 3rd Annual ACM Web Science Conference. the 3rd Annual ACM Web Science Conference. Evanston, Illinois. Web Science 2012; ACM Special Interest Group on Hypertext, Hypermedia, and Web. New York, NY: ACM, pp. 24–32.

Back, Mitja D.; Stopfer, Juliane M.; Vazire, Simine; Gaddis, Sam; Schmukle, Stefan C.; Egloff, Boris; Gosling, Samuel D. (2010): Facebook profiles reflect actual personality, not self-idealization. In *Psychological science* 21 (3), pp. 372–374. DOI: 10.1177/0956797609360756.

Bai, Shuotian; Zhu, Tingshao; Cheng, Li (2012): Big-Five Personality Prediction Based on User Behaviors at Social Network Sites, 4/21/2012. Available online at http://arxiv.org/pdf/1204.4809.

Barrik, Murray; Mount, Michael (1991): The Big Five personality dimensions and job performance. A meta-analysis. In *Personnel Psychology* 44 (1), pp. 1–26. DOI: 10.1111/j.1744-6570.1991.tb00688.x.

Bouchard, Thomas J.; McGue, Matt (2003): Genetic and environmental influences on human psychological differences. In *Journal of neurobiology* 54 (1), pp. 4–45. DOI: 10.1002/neu.10160.

Carney, Dana R.; Colvin, C. Randall; Hall, Judith A. (2007): A thin slice perspective on the accuracy of first impressions. In *Journal of Research in Personality* 41 (5), pp. 1054–1072. DOI: 10.1016/j.jrp.2007.01.004.

Fast, Lisa A.; Funder, David C. (2008): Personality as manifest in word use. Correlations with self-report, acquaintance report, and behavior. In *Journal of Personality and Social Psychology* 94 (2), pp. 334–346. DOI: 10.1037/0022-3514.94.2.334.

Fruyt, Filip de; McCrae, Robert R.; Szirmák, Zsófia; Nagy, János (2004): The Five-factor Personality Inventory as a measure of the Five-factor Model: Belgian, American, and Hungarian comparisons with the NEO-PI-R. In *Assessment* 11 (3), pp. 207–215. DOI: 10.1177/1073191104265800.

Funder, David C. (2012): Accurate Personality Judgment. In *Curr Dir Psychol Sci* 21 (3), pp. 177–182. DOI: 10.1177/0963721412445309.

Furnham, Adrian; Petrides, K.V; Jackson, Chris J.; Cotter, Tim (2002): Do personality factors predict job satisfaction? In *Personality and Individual Differences* 33 (8), pp. 1325–1342. DOI: 10.1016/S0191-8869(02)00016-8.

Goel, Sharad; Hofman, J. M.; Sirer, M. Irmak (2012): Who does what on the Web. Studying Web browsing behavior at scale. In : International Conference on Weblogs and Social Media, pp. 130–137.

Gosling, Samuel D.; Ko, Sei Jin; Mannarelli, Thomas; Morris, Margaret E. (2002): A room with a cue. Personality judgments based on offices and bedrooms. In *Journal of Personality and Social Psychology* 82 (3), pp. 379–398.

Grassegger, Hannes; Krogerus, Mikael (2016): Ich habe nur gezeigt, dass es die Bombe gibt. Der Psychologe Michal Kosinski hat eine Methode entwickelt, um Menschen anhand ihres Verhaltens auf Facebook minutiös zu analysieren. Und verhalf so Donald Trump mit zum Sieg. 48. Available online at https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt/, checked on 1/17/2018.

Guyon, Isabelle; Weston, Jason; Barnhill, Stephen; Vapnik, Vladimir (2002): Gene Selection for Cancer Classification using Support Vector Machines. In *Machine Learning* 46 (1/3), pp. 389–422. DOI: 10.1023/A:1012487302797.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2009): The elements of statistical learning. Data mining, inference, and prediction. 2. ed. New York NY: Springer (Springer series in statistics).

Hirsh, Jacob B.; Kang, Sonia K.; Bodenhausen, Galen V. (2012): Personalized persuasion. Tailoring persuasive appeals to recipients' personality traits. In *Psychological science* 23 (6), pp. 578–581. DOI: 10.1177/0956797611436349.

Jernigan, Carter; Mistree, Behram F.T. (2009): Gaydar. Facebook friendships expose sexual orientation. In *First Monday* 14 (10). DOI: 10.5210/fm.v14i10.2611.

John, Oliver P.; P. Naumann, Laura; J. Soto, Christopher (2008): Paradigm Shift to the Integrative Big Five Trait Taxonomy. In Oliver P. John, Richard W. Robins, Lawrence A. Pervin (Eds.): Handbook of personality. Theory and research. 3. ed. New York, NY: Guilford Press, pp. 114–117. Available online at https://www.ocf.berkeley.edu/~johnlab/pdfs/2008chapter.pdf, checked on 12/1/2017.

John, Oliver P.; Srivastava, Sanjay: The Big Five Tait Taxonomy. History Measurement, and Theoretical Perspectives.

Kosinski, Michal; Stillwell, David; Graepel, Thore (2013): Private traits and attributes are predictable from digital records of human behavior. In *Proceedings of the National Academy of Sciences of the United States of America* 110 (15), pp. 5802–5805. DOI: 10.1073/pnas.1218772110.

Kuhn, H. W.; Tucker, A. W. (1951): Nonlinear Programming. In. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability: The Regents of the University of California. Available online at https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200500249.

Mairesse, François; Walker, Marilyn A.; Mehl, Matthias R.; Moore, Roger K. (2007): Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In *J. Artif. Intell. Res.* 30, pp. 457–500.

Marcus, Bernd; Machilek, Franz; Schütz, Astrid (2006): Personality in cyberspace. Personal Web sites as media for personality expressions and impressions. In *Journal of Personality and Social Psychology* 90 (6), pp. 1014–1031. DOI: 10.1037/0022-3514.90.6.1014.

McCrae, Robert R.; Costa, Paul T.; Ostendorf, Fritz; Angleitner, Alois; Hřebíčková, Martina; Avia, Maria D. et al. (2000): Nature over nurture. Temperament, personality, and life span development. In *Journal of Personality and Social Psychology* 78 (1), pp. 173–186. DOI: 10.1037/0022-3514.78.1.173.

Noftle, Erik E.; Shaver, Phillip R. (2006): Attachment dimensions and the big five personality traits. Associations and comparative ability to predict relationship quality. In *Journal of Research in Personality* 40 (2), pp. 179–208. DOI: 10.1016/j.jrp.2004.11.003.

Oberlander, Jon; Nowson, Scott (2006): Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 627–634. Available online at http://www.aclweb.org/anthology/P06-2081.

Provost, Foster; Fawcett, Tom (2013): Data science for business. [what you need to know about data mining and data-analytic thinking]. 1. ed. Sebastopol Calif. u.a.: O'Reilly.

Quercia, D.; Kosinski, M.; Stillwell, D. J.; Crowcroft, J. (2011): Our Twitter profiles, our selves. Predicting personality with Twitter. In *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, pp. 180–185. Available online at https://www.cl.cam.ac.uk/~dq209/publications/quercia11twitter.pdf.

Rentfrow, Peter J.; Gosling, Samuel D. (2003): The do re mi's of everyday life. The structure and personality correlates of music preferences. In *Journal of Personality and Social Psychology* 84 (6), pp. 1236–1256.

Smola, Alex J.; Schölkopf, Bernhard (2004): A tutorial on support vector regression. In *Statistics and Computing* 14 (3), pp. 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88.

Soldz, Stephen; Vaillant, George E. (1999): The Big Five Personality Traits and the Life Course. A 45-Year Longitudinal Study. In *Journal of Research in Personality* 33 (2), pp. 208–232. DOI: 10.1006/jrpe.1999.2243.

Statista (2017): Most famous social network sites worldwide as of September 2017, ranked by number of active users (in millions). 2017 (Social Media & User-Generated Content). Available online at https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Statistisches Bundesamt (3/16/2015): 58,6 Millionen Internet-nutzer/-innen in Deutsch-land im 1. Quartal 2014. Wiesbaden, Auskunftsdienst Einkommen, Konsum, Lebensbedingungen,. PDF file. Available online at https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2015/03/PD15_098_63931.html, checked on 1/17/2018.

van Belle, Vanya; Lisboa, Paulo (2014): White box radial basis function classifiers with component selection for clinical prediction models. In *Artificial intelligence in medicine* 60 (1), pp. 53–64. DOI: 10.1016/j.artmed.2013.10.001.

Vapnik, Vladimir N. (2000): The Nature of Statistical Learning Theory. Second Edition. New York, NY: Springer (Statistics for Engineering and Information Science). Available online at http://dx.doi.org/10.1007/978-1-4757-3264-1.

Youyou, Wu; Kosinski, Michal; Stillwell, David (2015): Computer-based personality judgments are more accurate than those made by humans. In *Proceedings of the National Academy of Sciences of the United States of America* 112 (4), pp. 1036–1040. DOI: 10.1073/pnas.1418680112.

APPENDIX

Plots

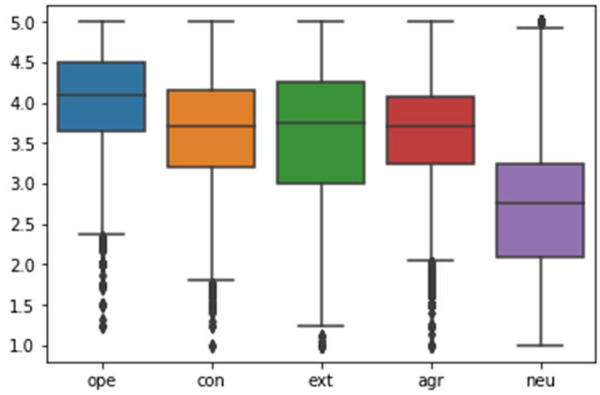


Figure 4: Target variables boxplot

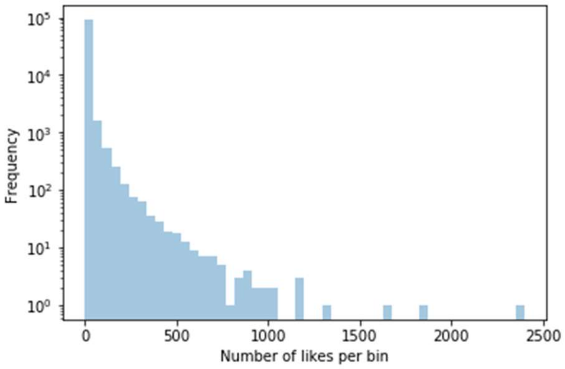


Figure 5 - Distribution of likes per page. The pages have been aggregated in 20 bins. X-axis shows the number of likes and the log-scaled y-axis the frequency.

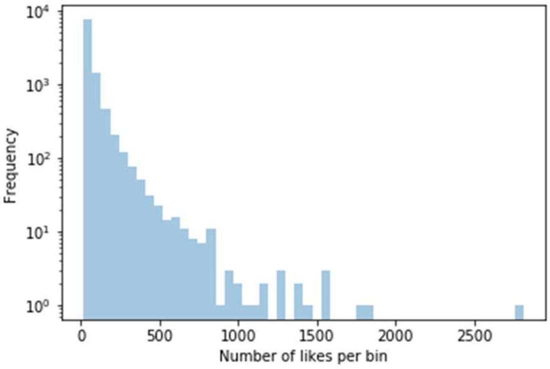


Figure 6 - Distribution of likes per user. The users have been aggregated in 20 bins. X-axis shows the number of likes and the log-scaled y-axis the frequency.

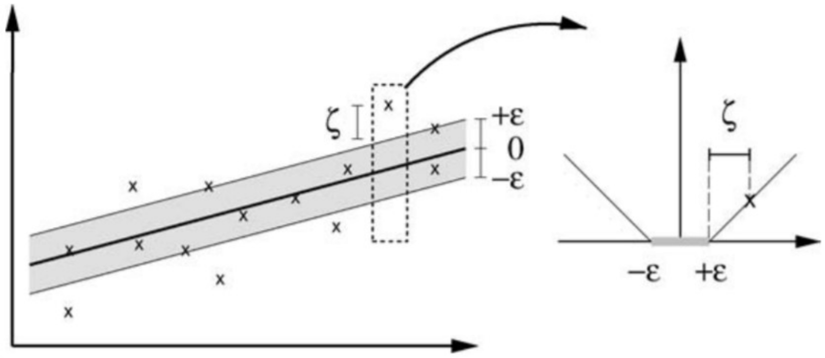


Figure 7 – from (Smola, Schölkopf 2004). ε-insensitive loss and slack variables. Just deviations outside the loss-insensitive margin contribute to the cost function in a linear fashion relating to C.

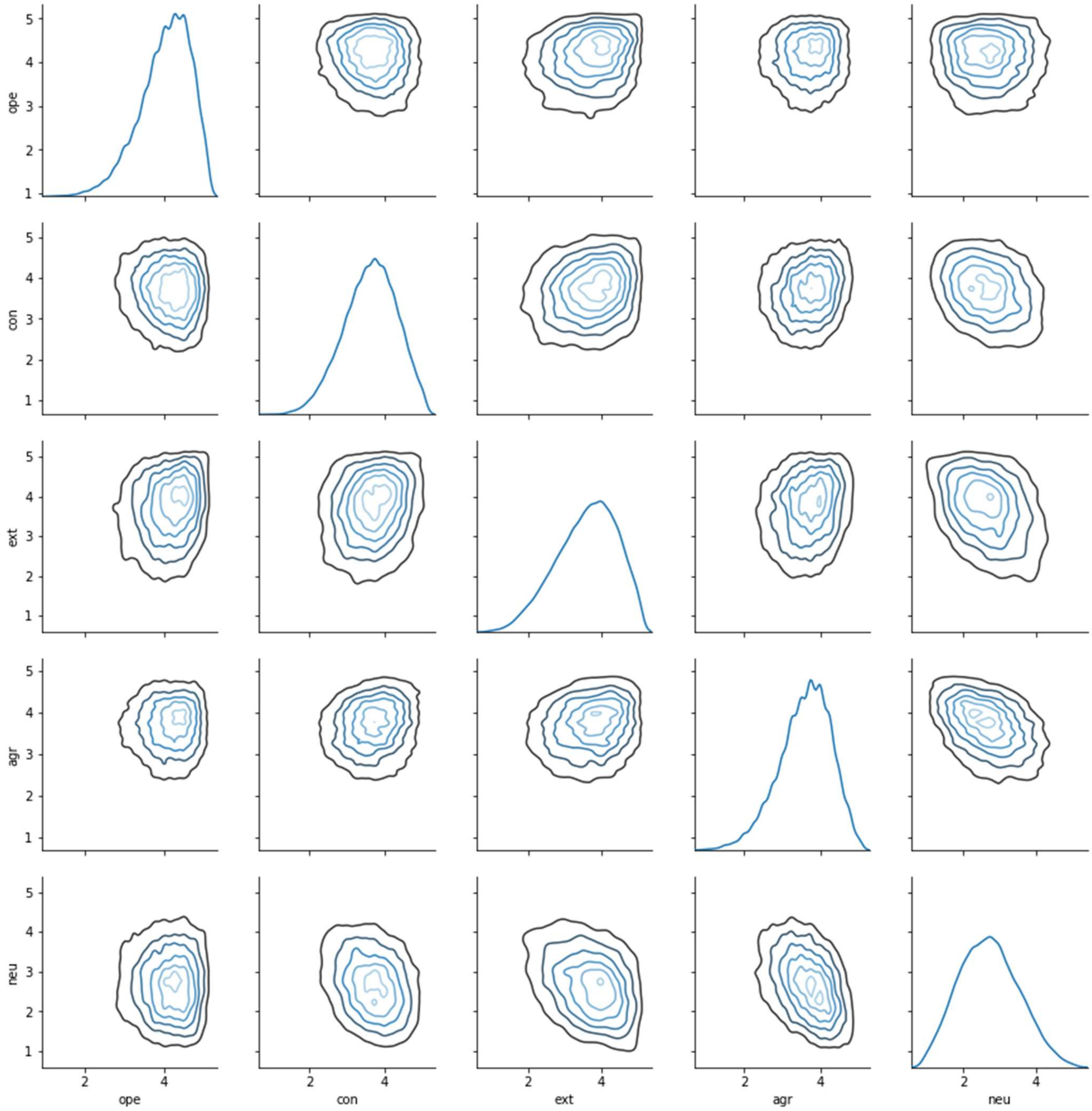


Figure 8 - Target variable (paired co-) density plots

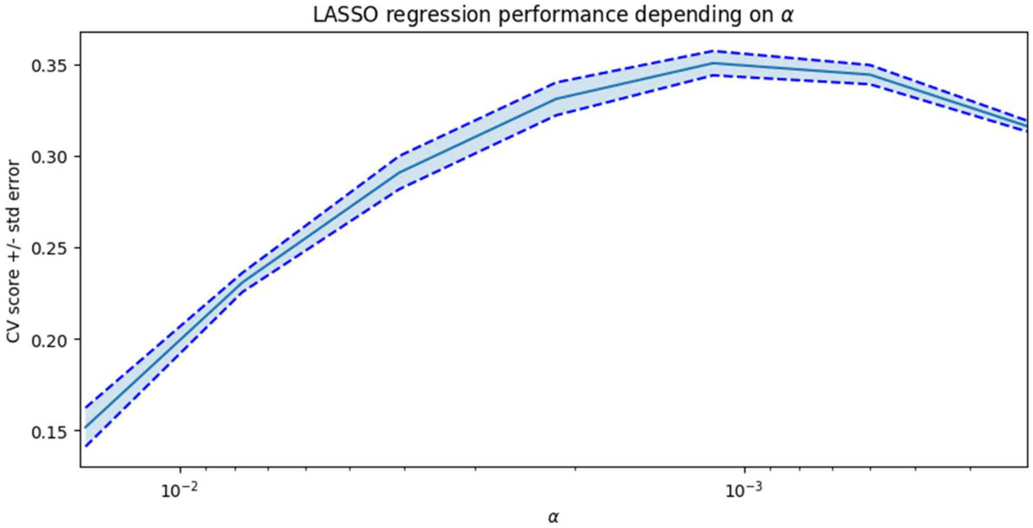


Figure 9 – LASSO regression performance depending on α (N=10.000) predicting extraversion (using 10-fold cross validation). Performance pattern and peak performance is generalizable at a level of PCC~0.35. The dottend line represents actual accuracy, the light blue area indicates the standard deviation.

EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich, dass

• die Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt wurden,

• alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht wurden,

• die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Lüneburg, 16.02.2018