Case 2 – Fundamentals of Data Science

Melania Berbatova , (11961279) , Michal Kacprzak (11871695) ,
Bob van de Mortel (11720239) , and Jiayang Zhuo (11611359)

1. Introduction

With the rise of global competition and more demanding customers, nowadays many organisations are seeking new ways to achieve competitive advantage. In the past, attempts were directed at quality management, restructuring and reengineering. However the rise of big data and web scraping techniques parallel to the increasing usage of new media has brought marketeers to a new frontier. The next step will be to utilize these techniques and platforms into the direction of outward orientation towards customers, to put an emphasis on the customer value for the organisation. Customers value is defined by "customer's preference for and evaluation of product attributes, attribute performances, and consequences arising from use that facilitate (or block) achieving the customer's goals and purposes in use Situations" (Robert B. Woodruff).

Customers goals may vary drastically among customers from different age groups, gender, location, income and cultural background, but there is one common goal that people around globe have: to achieve happiness. Understanding what makes people happy, tailoring to that sentiment, and associating their brand with these circumstances can help businesses design better products and services, that respond to customer's purposes and lead to more satisfied and loyal customers.

With the rise of popularity of social media and with the new methodologies available for processing big amounts of structured and unstructured data, marketing researchers are provided with new means to gain insights about customers preferences and to predict customers behaviour (Wedel, Kannan, 2016).

2. Research question

The purpose of our current project is to develop a model to predict customers' levels of happiness such that we can later correlate this with their circumstances in order to specify target groups and tailor our products.

Firstly, we should answer the question how we can operationalize the concept of happiness. In general, happiness is a vague idea, but there are methods to represent someone's happiness with a numerical value. The first step that we took is to adopt the broad definition of happiness that encompasses all elements of utility. The method that we will use is the PERMA model. PERMA score consist of five factors: Positive Emotion, Engagement, Relationships, Meaning and Accomplishments. Individual PERMA scores are evaluated by conducting a survey with questions with possible answers in set of integers in the range from 1 to 10 referring to all of the five components and and taking the mean as a single score.

Secondly, we consider how we can find data and build a model that can be used to predict these PERMA scores. We can again conduct a survey, but there are several issues to be considered. Surveys are time and budget-consuming and we can not guarantee accurate results. Since we have an exposure of all kind of publicly available personal data on the web, we can use it to train models for predicting happiness. For that purpose, we will use publicly available photos of Instagram users and try to gain information about factors that give them utility or happiness as addressed by the PERMA scores.

3. Data processing

In order to utilize the Instagram images to predict the happiness of users, several image analysis algorithms were applied on a sample of streamed Instagram images, posted by the people who have taken part in the happiness survey. Features are extracted from the images and distributed to 5 datasets, including the ANP, Face, Object label, Image info and Metrics, which can be defined as our predicting datasets. Only a small part of features are explicit and represented with a single numerical value. For the other features, for example these obtained from the image recognition algorithm (such as ANP label, sentiment label and emotional score), we take into consideration the confidence score accompanying them.

In the initial data, there was an error in calculating PERMA scores. Instead of the mean of every component, P, E, R, M, A, H, N_emo and PERMA were formed by just taking the first value. We recalculated the scores and used the new values in our prediction methods.

Data transformation, feature engineering and aggregation

The datasets with potential predictors contain huge amounts of non numerical data, which we cannot directly use for the prediction model, either regression or classifier. We transformed them using the confidence levels given. In order to establish a proper model, we first aggregated the five predicting datasets on image level and then combine them with the survey data by the user id.

For our model, we selected only images posted 3 months before and after the survey, as we would expect them to reasonably include data on the PERMA scores captured by the survey results. This is done under the assumption that people's happiness changes over the course of time and there would be a tradeoff between having more images to support our model and accuracy with regards to our predictor variable which is a point estimate at a specific moment. We also chose only users with 5 or more pictures, as having not enough data for a user can block us from having accurate results and introduces outliers in the data.

Some important categorical features like ANP-label, emotion label, multiple face features, data amz labels, and selected Instagram filters, would seem theoretically relevant for the prediction.

However, instead of using all of the labels, we select the ones that are most representative for the images in order to be able to aggregate them in a meaningful way to image and user level and derive conclusions of trends within our limited dataset. Therefore, we select the 80 most occurring labels and decide which of these is most applicable to the users' images by taking the largest confidence score and frequency across multiple images. The other labels are dropped since their occurrence is very rare and based on singular cases, thus not fit for a predictive model.

Afterwards all other labels, such as the data amz labels and the face recognition labels will be implemented by one hot encode to transfer the categorical data into binary data and then weighted by the confidence as float data type. In the end, each image is aggregated by the sum of weighted value on each label by image ID and then merged with the survey data by user ID. In addition, the image filter is selected by the most used filter on every user, excluding default mode "normal", because the normal filter is the default setting and does not necessarily capture any information on the user's mood. These methods ensure that we maintain the most comprehensive overview of the users' image data when aggregating to the user level.

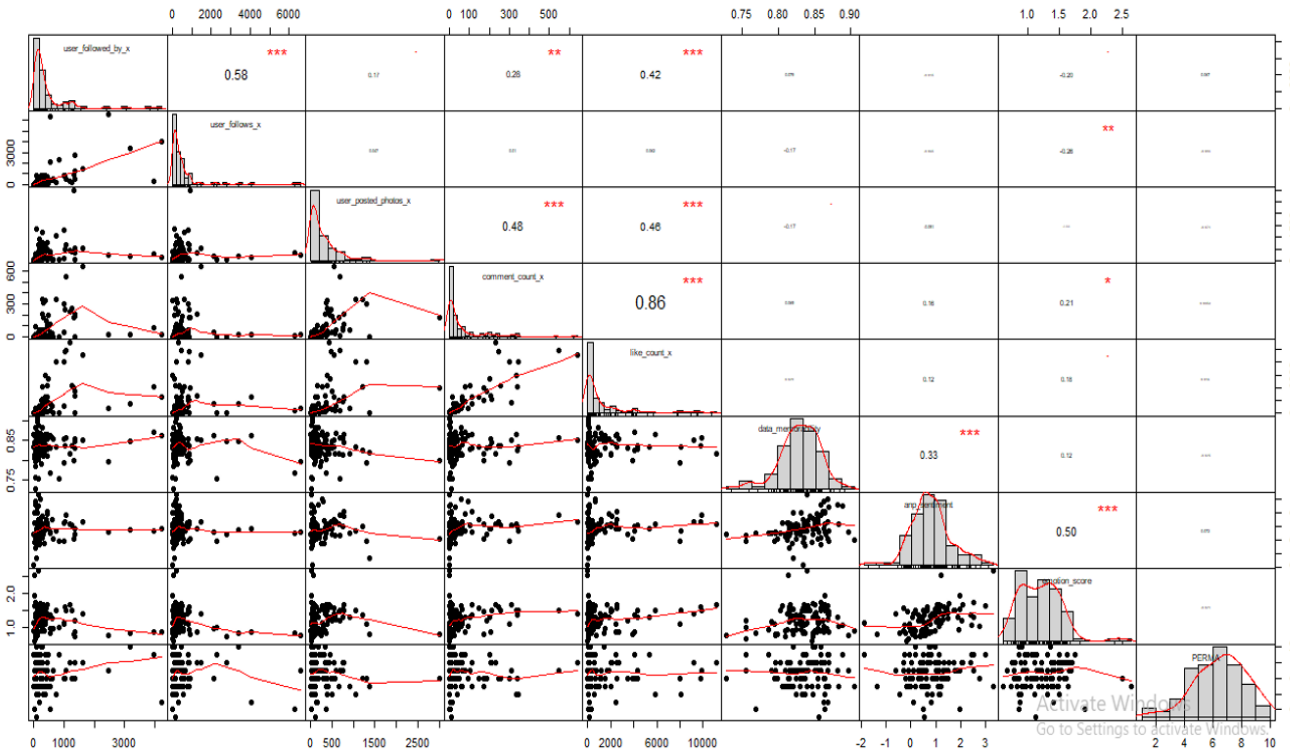

Figure 1: Exploration of variables' distribution and correlation

After preparing all the data, and splitting our data in a train and test dataset (80% and 20% of observations respectively), we ran a Spearman rank correlation analysis on all the variables in our training dataset and investigated their distributions. As can be seen in the subset of variables shown in figure 1, this displayed the need for negative binomial transformation of some variables. A log transformation was not in order in many cases because there was a meaningful zero value in our count data that we did not want to ignore.

Among other observations, we saw that number of followers and number of followed users are highly correlated variables, as well as number of comments, number of likes per picture and number of pictures. Instead of using them directly, we transformed them by adding new variables for follower: followers and follows ratio and average number of comments and likes per picture. We also group emotional labels in 2 groups, adding 2 more variables for positive emotions and negative ones.

4. Prediction methods

The large output of the Spearman rank correlation over the many variables included in our dataset showed many collinear as well as many largely irrelevant variables. If we would use all these variables in our model, then we would risk overfitting the data and introducing unnecessary complexity. Therefore, we decided to use principal component analysis in order to help us reduce the number of predictor variables in the model. As figure 2 shows, most of the variance in the dataset can be explained by using just five components. We can thus reduce our many predictor variables to five component classes that we use to predict the PERMA scores.

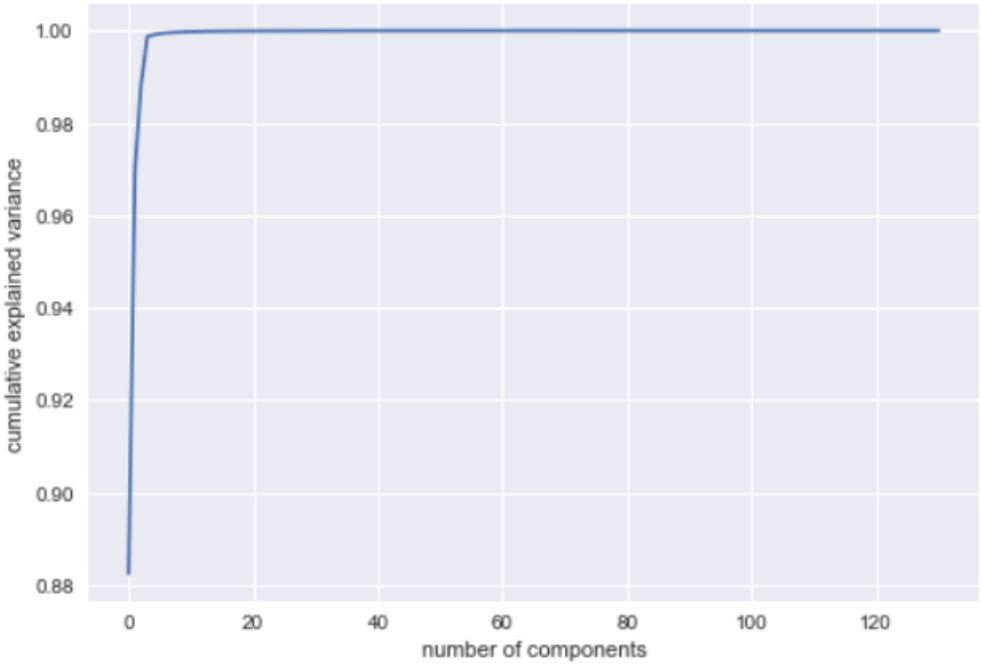

Figure 2: Explained variance

5. Results

We ran the five components identified in the principal component analysis in a regression model to predict the PERMA scores in our training dataset and cross-validated our in-sample predictions with five splits of the training and testing dataset. This turned out to be necessary because of the small amount of observations that was included in our datasets. We achieved the following mean squared errors and variance scores over our iterations as shown in Table 1:

	Split 1	Split 2	Split 3	Split 4	Split 5
Mean squared error	-2.29	-1.64	-3.13	-3.29	-0.88
Variance score	-0.96	-0.84	-0.15	-0.65	0.30

Table 1: Results from five iterations of our model with different splits of the data to cross validate our in sample predictions of the model

Even though none of these model iterations really gave very good predictions, we decided to build upon the last iteration of our model and run the model against the actual values of our test dataset. Hereby we would assume that increasing the number of observations in production versions of such a model would converge towards smaller errors as per the law of large numbers. The results are shown in table 2 and plotted in figure 3

Mean squared error	1.67
Variance score	0.28

Table 2: Results of our model predictions with regards to our out-of-sample observations.

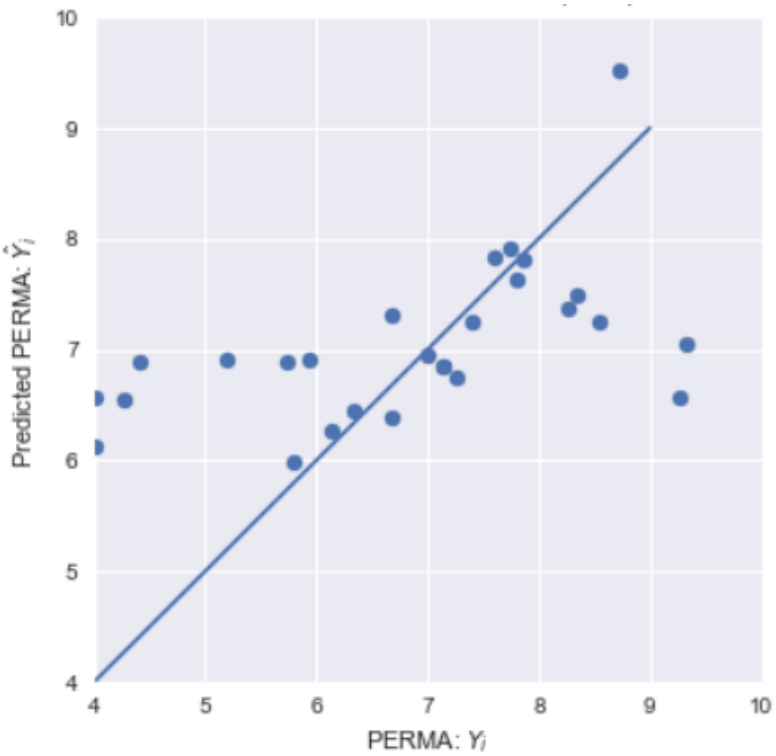

Figure 3: Observed PERMA scores vs predicted PERMA scores from test dataset. The line represents the ideal trend of model predictions.

As can be seen from the variance score of 0.28 we do explain a portion of the variance that occurs in the dataset, though we incur a relatively large mean squared error of 1.67 on our 10-point PERMA scale. Figure 3 shows our predictions of the PERMA scores for the observations in the test dataset plotted against their actual PERMA scores. We observe that we see a less steep trend in our own predictions than in the ideal true predictions as shown by the blue trendline. These results give an indication that there is a fruitful basis in Instagram data to predict PERMA scores, but that the dataset should be expanded in terms of observations in order to improve the predictive power of the model.

6. Ethical issues

We also reviewed this project from an ethical perspective, since during our investigation several ethical concerns came to our attention. The approach that we used to tackle these problems is the Iterative Reflexivity Methodology (Zevenbergen 2016). It consists of three steps which need to be repeated until the final model of the project is reached and justified:

1. Think about the context of the project - benefits, methods, technology.
2. Draw out ethical tensions.
3. Suggest alternatives which could reduce risks of harm.

During this IRM process, we encountered several ethical issues related to our project. First of all, the data set includes excessive amounts of data as well as excessive data fields which involve personal information. We excluded these from consideration and also did not retain the raw data files that contained this information.

Secondly, it is not clear if the data subjects gave consent to using their data for scientific purposes. Since we were supplied with the survey dataset after it had been collected, we could not reasonably retroactively request permission for the use of the data from the subjects. However, we do want to address that this would be a necessary condition in any further research that should be tackled.

Thirdly, we are aware of the function creep that exists in the implementations of this technique. It could be exploited by health insurance companies and employers to identify customers or employers at risk of getting a depression and reducing costs by firing these people before they disclose this information. Our solution to mitigate this risk is to not be explicit about the code or the components used in our final model, in order to prevent direct usability of our methods without our consent.

Becoming aware of these issues, we have mitigated some of the ethical tensions listed above and could carry out the project in relative harmony with ethical guidelines. Further research should take these into account in order to introduce precautions to reduce the threats of this data research to all stakeholders involved - be it society at large or surveyed individuals.

7. Conclusion and suggestions for future research

The model that we come up with is not very accurate in predicting individual PERMA scores. However, it does show us that there is a potential predictive power of Instagram data that can be obtained to predict the trends in these scores.

The main factor that would contribute to the success of further investigation of these predictions is the expansion of the amount of observations in the dataset. Only 159 out of 331 users taking part in the survey had public profiles and not all of them had a sufficient number of photos to make meaningful predictions. With more data, one could make more accurate predictions.

There are many theoretical justifications for including certain predictors in a model. For example studies on environment and behaviour show that nature and animals make people happy (Zelenski and Nisbet 2012), and in a study of Instagram photos predicting depression, authors found out that colors of pictures and average number of people on the pictures are good markers of people's depression and are good predictors of happiness as well (Reece, Danforth, 2017).

However, because of the Principal Component Analysis that we conducted, we eradicated some of the factors that have theoretical support in relation to depression and happiness scores. This is a choice in order to limit the risk of overfitting our model since we have so many variables.

The support of the literature however, warrants conducting a separate investigation that uses a slimmer set of theoretically relevant predictors over a larger amount of observations in order to achieve meaningful results.

In conclusion, a larger survey dataset (with informed consent of the subjects) and more resources in terms of image analysis with regards to hue and saturation would have given us the opportunity to construct a model with more accurate individual predictions. However, the results do show already that information regarding happiness (approximated by PERMA scores) is able to be obtained from Instagram data to an extend.

This provides perspective for marketeers willing to use such data to present their products to audiences that fit to certain consumer clusters, or they could develop a model to analyse the satisfaction of an existing consumer base when selecting users based on tags or recognized objects related to their brand.

References

Pascha, Mariana. 2017. "The PERMA Model: Your Scientific Theory of Happiness." February 24, 2017. https://positivepsychologyprogram.com/perma-model/.

Reece, Andrew G., and Christopher M. Danforth. 2017. "Instagram Photos Reveal Predictive Markers of Depression." *EPJ Data Science* 6 (1):15. https://doi.org/10.1140/epjds/s13688-017-0110-z.

Wedel, Michel, and P.k. Kannan. 2016. "Marketing Analytics for Data-Rich Environments." *Journal of Marketing* 80 (6):97–121. https://doi.org/10.1509/jm.15.0413.

Woodruff, Robert B. 1997. "Customer Value: The next Source for Competitive Advantage." *Journal of the Academy of Marketing Science* 25 (2):139. https://doi.org/10.1007/BF02894350.

Zelenski, John M., and Elizabeth K. Nisbet. 2014. "Happiness and Feeling Connected: The Distinct Role of Nature Relatedness." *Environment and Behavior* 46 (1):3–23. https://doi.org/10.1177/0013916512451901.

Zevenbergen, B. 2016. "Networked Systems Ethics." 2016. http://networkedsystemsethics.net/index.php?title=Networked_Systems_Ethics_-_Guidelines.