Assignment 2: Business (Instagram)

Group 4 - Fundamentals of Data Science
Authors: Milos Dragojevic (10853456), Joris van der Vorst (10210717), Tycho Atsma (10791779), Andrew Paterson (12777153), Antonio Javier Samaniego Jurado (12799505).

Abstract

Instagram is a platform where millions of people share visual content with each other based on their current mood or emotional state. This project makes use of a regression model that predicts individual well-being state as expressed in the PERMA score by applying data analysis and feature engineering techniques on Instagram visual and user data.

Introduction

In this project, a combination of data analysis, feature engineering and regression techniques are performed on a dataset of visual features extracted from Instagram data (images) and survey data from users, which includes the PERMA score, to predict individual well-being. The performed feature engineering consists in the manipulation of strategic features and later combination based on two methods: standard user level aggregation and time-penalized user level aggregation. Using the method along with different regression techniques, we trained a model that is able to predict the well-being state as in PERMA score with reasonable accuracy, taking RMSE and R^2 as the reference error metrics. A research question on what is the nature of relationship between Instagram posts and well-being is both defined and answered in detail, along with a set of insights.

Methodology

The analysis was performed using five different datasets. Four of these datasets contained data related to emotional semantic constructs (ANP), facial features and emotions, image metadata, and user data. These all relate to instagram activity and image content. The fifth database contains survey data with expresses the well-being of these users in a PERMA score.

Methodology: Data aggregation

The four datasets mentioned in the previous section describe the user Instagram activity in one or more images, where each image is described through one or more features in one or more rows. This dimensionality differs from the survey dataset, where each entry is independent. Therefore, the data has

to be reduced so that each entry in the survey dataset can be compared to a single set of data describing that user instagram activity and image content. Related work performed by Schwartz et al. (2016) shows how a *cascading* model can be used to identify and approach such a problem.
The following section explains how this problem is tackled through data aggregation.

Methodology: Feature aggregation

This section describes how the *face* dataset is aggregated. This dataset described facial features, like *face_beard*, based on an image. There can be one or more entries per image, based on the number of faces that have been recognized in an image. Most facial features are described in multiple columns. For example, facial hair is described through four columns: *face_beard, face_beard_confidence, face_mustache, and face_mustace_confidence*.
To simplify and prevent unexpected correlations between parameters such as confidence levels of facial features like *face_beard_confidence* and *smile_confidence*, all values that describe a single feature, like facial hair, have been merged into a single numerical value. This is done in several steps.
First, there are several groups with just two values: some value *x* and the confidence level of value *x* expressed in *y*. These have been reduced to a single value, where the confidence *y* determines the outcome of that value.
Second, the *emotion* group has been one-hot-encoded since it is a categorical variable. The goal is to predict a PERMA score, which is based on the emotional state of a human being. Therefore, it was decided that these could have a significant influence on the target result.
Finally, after grouping the variables, all entries that point to the same image have been merged together into a single entry. This is done through either calculating the mean, mode, or sum of those values.

User aggregation

After all the features of an image were combined into a single row, all images were grouped by user and all numerical values where aggregated into a single row per user. This was done in two ways:
The first method treated every image of the user as equally relevant and, for each numerical feature, (including the hot encoded categorical values) the mean over all images was recorded.
The second method included a time penalty for older images to test whether more recent, current pictures, reflect the well-being state of a user more accurately (e.g. a user could be depressed a month ago but not today). This was achieved by creating a “time score” that penalizes image IDs based on age. The $score_i$ for image i is given by the proportion of elapsed time $\left|t_i - t_L\right|$ (where t_i is the time image i was posted, and t_L is the time the latest image in the dataset was posted) across the total range of posting times $T = \left|t_O - t_L\right|$, where t_i is the time the oldest image in the dataset was posted. Thus:

$$score_i = 1 - \frac{|t_i - t_L|}{T}, \text{ where } 0 \leq score_i \leq 1$$

The resulting time penalization scores for this dataset are shown in *figure 1*.

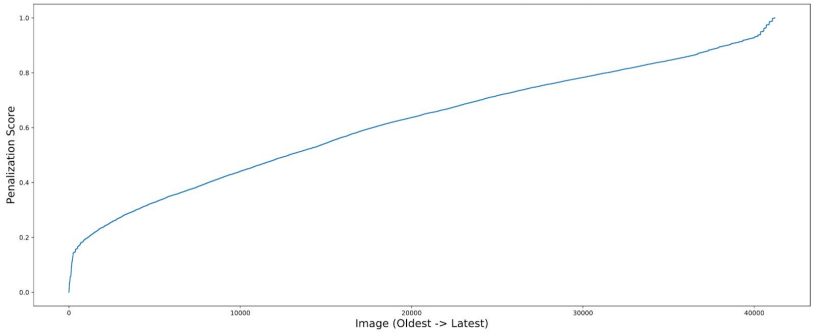

Fig 1. - Penalization Score vs. Image Posted (Oldest -> Latest)

The above figure shows that recent image IDs (i.e. row vectors of each *image_id*) are multiplied by a progressively higher score to then be aggregated by user, and vice versa. This gives more weight to recent images and models that current pictures reflect the emotional state of a user more accurately. We also added an extra column *elapsed_time* $= |t_i - t_L|$ to the image data.

All non-numerical values that were unique to each user (links, usernames, etc) were dropped from the analysis dataset.

Missing Data

The final dataset was checked for missing data. One PERMA score was not calculated due to a missing value for the happiness question and 14 users had no faces in any of their pictures and therefore no facial data. The missing happiness question was imputed by taking the median of the values in the dataset and all the missing facial feature data was set to 0 (since none of the features was detected in the images of those users).

Transformations

The histograms for all features were checked and "user_followed_by" , "user_follows" and "user_posted_photos" were log10 transformed to more closely resemble a normal distribution.

Using a correlation matrix all features were checked for pairwise correlations. Three combinations', were found to have a $R^2 > 0{,}75$. These features combination were compared with the target value ‘PERMA’ and the ones that correlated the least were dropped from the dataset.

Validation set

Ten percent of the dataset was randomly selected and set apart for validation in order to test final models on data unused for training or hyperparameter tuning purposes.

Multiple linear regression

A multiple linear regression is used to model the relationship between multiple explanatory variables and a response variable by fitting a linear equation to the observed data. Thus using a variety of parameters we are trying to estimate what the PERMA Score would be.

LightGBM, a gradient boosting framework that uses tree based learning algorithms, has also been tested, choosing the application for regression.

Forward Sequential Feature Selection

In order to select a subset of features for our linear regression model, we used the MLxtend Forward Sequential Feature Selector by Raschka, (2018). This python package iterates through the set of possible features, selects the feature with the highest possible performance metric (in our case R^2) and adds this to the feature subset. Subsequently all other possible features are iterated trough and that performs the best in combination with the existing subset is added to the subset as well. This process is repeated until the subset with the best combination of features is returned.

Compound Linear Regression Model

The PERMA score is calculated by taking the mean of 6 factors: Positive Emotion, Engagement, Relationships, Meaning, Accomplishment and Happiness. Because these factors may represent themselves quite differently in the data, a compound model was created. A forward feature selection method was used to train 6 different linear regression models, one for each factor. The mean of the results was used to estimate a PERMA score from the individual predictions.

Results and Discussion

Full linear regression model:

Using a linear regression model with all features resulted in an R^2 of -10.927 and RSME of 5.246 for the standard user aggregation dataset and an R^2 of -7.126 and RSME of 4.330 for the time-penalized user aggregation dataset.

Best possible features:

Performance of the linear regression models related to the number of included figures can be seen in *figure 2* and *figure 3*. Using the Forward Sequential Feature Selector a subset of 10 features is chosen for the linear model of the standard user aggregation dataset as well as 13 features for the time-penalized user aggregation.

These models resulted in an R^2 of 0.156 and RSME of 1.187 for the standard user aggregation dataset and an R^2 of 0.259 and RSME of 1.112 for the time-penalized user aggregation dataset.

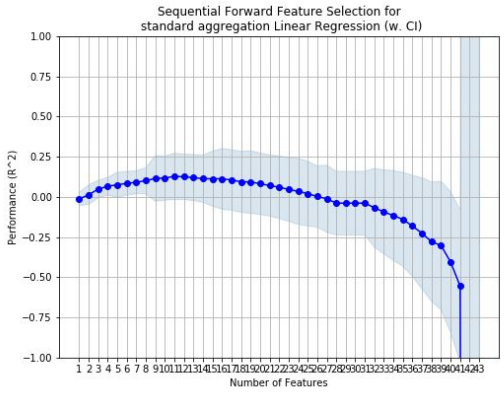

Fig. 2: Standard feature selection

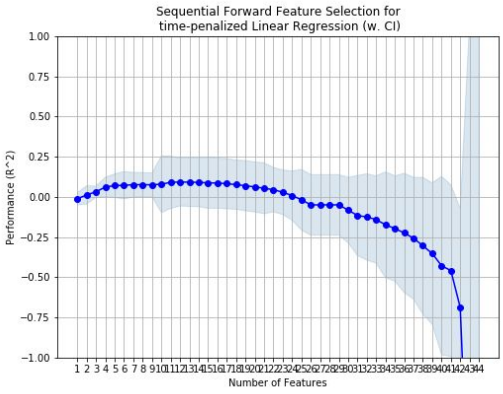

Fig. 3: time-penalized feature selection

Compound model:

The predictions using a combination of six linear models (one for each of the PERMA factors) resulted in an R^2 of 0.293 and RSME of 1.309 for the standard user aggregation dataset and an R^2 of 0.235 and RSME of 1.361 for the time-penalized user aggregation dataset.

Testing of the cleaned dataset with SKLearn Linear Regression and Support Vector Machine regressors

Once the provided dataset had been sufficiently reworked to allow regression testing an initial test was performed with two regressors provided in the SKLearn package: Linear Regression (LR) and Support Vector Machine (SVM) with a linear kernel. Train / test split for the regressors was arbitrarily chosen while testing progressed. For the images below the split was: LR: 70:30, SVM:90:10. The Random State in both cases was 42.

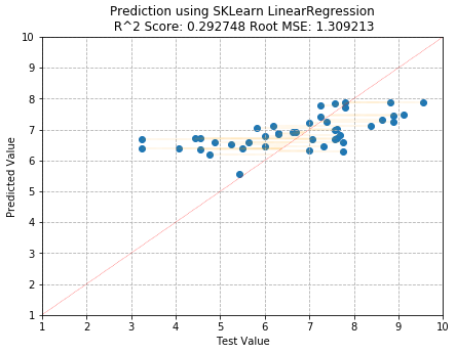

Fig. 4 - Prediction using LR

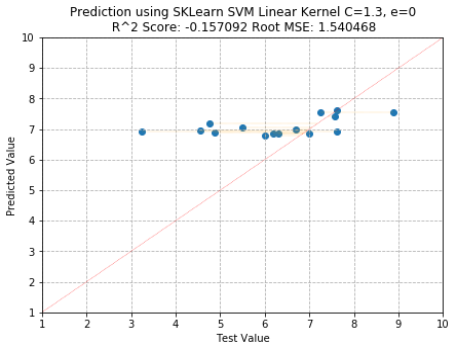

Fig. 5- Prediction using SVM

As can be noted from the images the predictions are similar. Differences in predicted values, R^2 and RMSE scores can be explained by the difference in size of the testing and training sets: LR had 66% more data available compared to the SVM while this test was performed.

Observations on R squared (R^2) and Root Mean Squared Error (RMSE) scores

Further evaluation was done as observed R^2 and RMSE scores on the entire dataset were perceived to show values denoting severe lack of fit. On the testing set, changes to the Random State resulted in comparatively large variations in these statistics. A test was performed where R^2 and RMSE were computed on results that were generated by data that was split with a randomly chosen value for the Random State of the Test/Train split method provided by SKLearn.

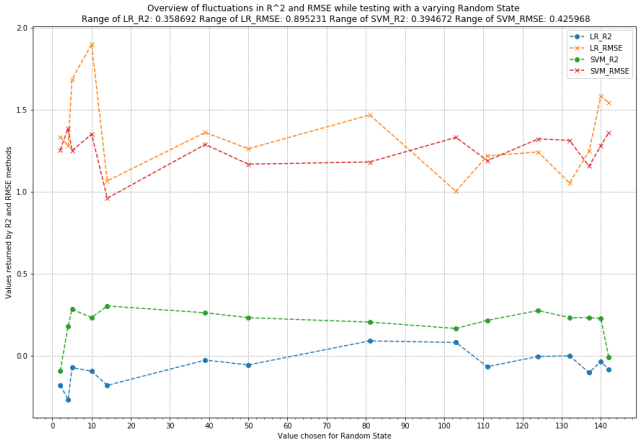

Fig. 6- Overview of Fluctuations in R^2 and RMSE

In the case of *figure 6*, the randomly chosen values for Random State were [2, 4, 5, 10, 14, 39, 50, 81, 103, 111, 124, 132, 137, 140, 142]. Fluctuations in the R^2 and RMSE were observed as the data was split with the values provided. A table was created providing statistics on the data as well as the range by which the statistic fluctuates.

	LR_R2	LR_RMSE	SVM_R2	SVM_RMSE
mean	-0.065729	1.350044	0.197565	1.253193
std	0.094989	0.247673	0.107401	0.109012
min	-0.267226	1.003835	-0.090573	0.960756
max	0.091466	1.899066	0.304099	1.386725
range	0.358692	0.895231	0.394672	0.425968

Table 1 - Mean, StdDev, Min, Max and range values for R2 and RMSE

The large range of the R^2 and RMSE statistics as noted from observations of the changes to these values, when experimenting with randomly chosen Random State values, implies a similar spread in the values predicted from the test dataset. This is reflected in *figure 4* and *figure 5* as generated during initial testing. Time penalties were applied on a copy of the dataset to verify if the notion of emotional state changing over time would be reflected in the prediction. This was indeed visible in a drift in predicted values. It was noted that R^2 and RMSE appeared to improve under this drifting.

Validation of models and methods

In order to get a sense whether our models will perform well all models were tested against the 10% of validation data that was not used for model training or hyperparameter tuning. The performance of all models for the test and validation dataset can be found in *table 2*. The data from the time-penalized user aggregation performed the best across all models. The compound linear regression model had the highest R^2 in both the test en validation datasets and the LightGBM had the lowest RMSE in the test dataset, but not in the validation dataset.

Method	**Model**	**Training**		**Validation**	
		R²	**RMSE**	**R²**	**RMSE**
Standard user aggregation.	LR all	-10.927	5.246	-1.645	2.089
	LR best features	0.156	1.187	-0.814	1.730
	LR compound	0.293	1.309	-1.463	2.016
	SVM	-0.157	1.540	-0.378	1.508
	LightGBM	*N/A*	0.8474	*N/A*	1.5303
Time-penalized user aggregation	LR all	-7.126	4.330	-1.294	2.149
	LR best features	0.259	1.112	0.311	1.177
	LR compound	0.235	1.361	0.326	1.164
	SVM	-0.179	1.555	0.083	1.358
	LightGBM	*N/A*	0.8741	*N/A*	1.5695

Table 2 - Overview of method, model, training and validation scores

Conclusion

This research was performed with the purpose of examining the link between user Instagram activity and the content of the images and their well-being.To that aim, a combination of data analysis, feature engineering and regression techniques was performed. Following a description of the data and methodology used throughout the study were presented. Based upon the data and the results, a hypothesis was formulated. The hypothesis examines the nature of relationship between Instagram posts and well-being. The results show that the compound linear regression model is the model which is the best able to explain the PERMA Score . Furthermore time-penalized user aggregation performed the best across all models.

One of the limitations of this study is the lack of available data. A more affluent dataset could have led to better prediction results. This is supported by the R^2 and the RMSE fluctuations found in the study.

Future research could focus itself on the reasons why well-being could fluctuate over time which could lead to more recent pictures being able to better reflect the current state of the user.

References

Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., & Kosinski, M. (2016). Predicting individual well-being through the language of social media. *Biocomputing 2016: Proceedings of the Pacific Symposium* (pp. 516-527).

Raschka, (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Journal of Open Source Software, 3(24), 638, https://doi.org/10.21105/joss.00638