STAKEHOLDER REPORT - #TAGSFORLIKES

This report is written by Marie-Louise Christensen, Pernille Franzen, Nina Danielsen & Lisa Nilsen

Instagram has become one of the most influential growing platforms in promoting brands and selling products, with more than 1 billion active users per month and 500 million daily active users. Moreover, 90 % of the 100 top brands in the world has an Instagram account which illustrate how important the media is in today's marketing (Lyfe Marketing, 2019). One-third of instagram's users have made an online purchase because of an advertisement they saw on Instagram and they are 70 % more likely to buy a product than people without an Instagram account (Lyfe Marketing, 2019). Therefore It's more important than ever for influencers, who promote these products, to know which content to post and when to do so, in order to get the most engagement on posts. When it comes to promotion, different kind of sources are referring to hashtags as one of the most popular methods in creating targeted campaigns. The use of relevant hashtags can help influencers create greater visibility for promotion of products and can make it easier for them to reach the target audience. According to Tweetangels, "when it comes to promotion, think of hashtags as the 'word-of-mouth' method." (Kathryn, 2019). In other words, when influencers are using hashtags to promote a product, service and event, then their followers tend to use that same hashtag, and then their followers will use it, which will create a snowball effect and the product will be promoted widespread (Kathryn, 2019). Therefore Instagram is a obvious choice for organisations to promote their brand. Knowledge of engaging content by using hashtags will benefit the organisations who choose to collaborate with an popular influencer because their brand will get noticed, and it can increase their sales rate. Furthermore, the influencers are measured by the likes and engagement they receive from their pictures. The more attention a post gets can influence the profit which the influencers will earn from the companies they are promoting. If the organisations are satisfied with the collaboration this can become the foundation for a longer-lasting partnership. Last but not least, knowing what indicates more likes and activity on a picture is relevant for upcoming influencers, who wants to increase their followers and to be better at targeting the right audience at the right time.

PURPOSE OF PROJECT

The aim of this project is to make a deep learning model that predicts if the number of likes are above or under the average based on hashtags. The project is targeted influencers who want to achieve more success and organisations who wants to use influencers to promote their brand. Influencers can use this model, to see which hashtags to use, to get a higher number of likes. Also, organisations can use it to see which influencers gets the most likes and choose them based on this. This can improve the promotion of a brand or a product.

INFO ABOUT THE DATASET

The chosen dataset is developed in 2017 and is called “instagram-like-predictor” which is obtained directly from github: https://github.com/gvsi/instagram-like-predictor.
The dataset consist of 2000 + different influencers on instagram. From each influencer the 17 most recent pictures has been chosen which add up to 16.538 observations. An observation consists of an image in JPG format, number of likes, number of comments, timestamp, description text. The number of likes and comments indicates the popularity of a post.

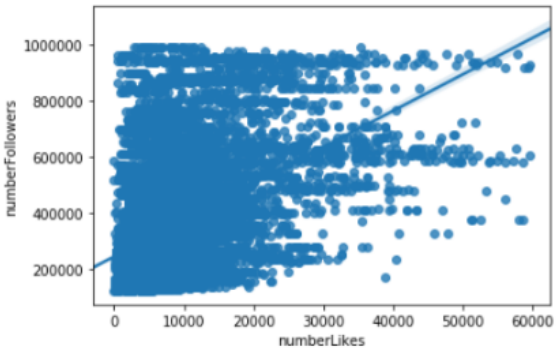
PREPROCESSING

Data exploration

First step is to explore the data to get an overview of all the interconnections between the chosen variables. This step will give us insight in which variables that is exciting to work further with. We want to compare four variables to see if there is a link between the number of likes you get and how many followers you have, what day of the week you post a picture and how many people you are following.

Likes and followers

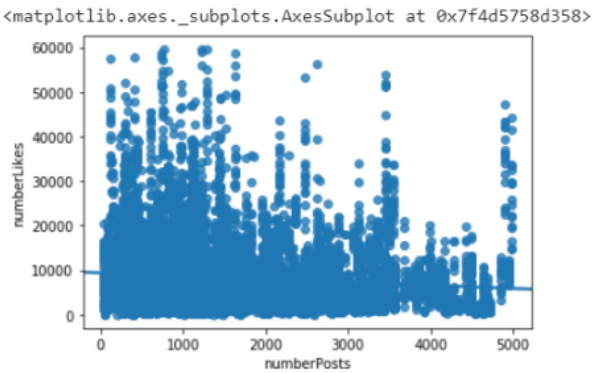
The first link we look at is between the number of likes you get and how many followers you have. The picture below visualize this causation.



From this we can conclude that more followers does not necessarily equals more likes, since those with 400.000 followers also receives up to 60.000 likes as the profiles with 100.000 followers do

Likes and posts

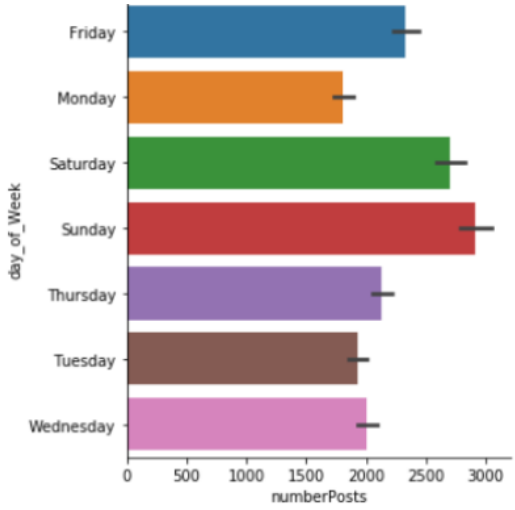
Next we will try to find out if the number of likes you get has something to do with how many people you are following.



From this we can conclude that posting a lot does not necessarily give you more likes.
On the other hand, we see that those with over 3500 posts in generals receives less likes.

Posts and day of the week

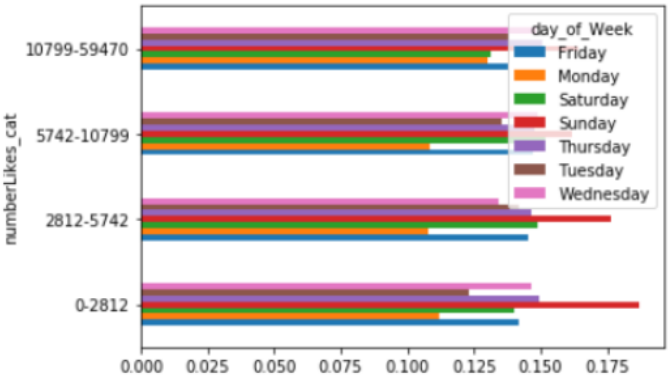
Next we will try to find out if the number of likes you get has something to do with how many people you are following.



Likes and day of the week

The table below shows that there is a higher engagement (likes) on sundays, thursdays and wednesdays. The day of the week with the lowest engagement is mondays. Since there is a higher engagement on sundays and thursdays, this would be a good time to post a picture to get the highest number of likes.

day_of_Week	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
numberLikes_cat							
0-2812	0.14	0.11	0.14	0.19	0.15	0.12	0.15
2812-5742	0.15	0.11	0.15	0.18	0.15	0.14	0.13
5742-10799	0.15	0.11	0.15	0.16	0.15	0.14	0.15
10799-59470	0.14	0.13	0.13	0.16	0.15	0.14	0.15



Sub-conclusion

Posting often and having the most followers does not equal more engagement it self, which may indicate the importance of the content of the post (i.e. image, hashtag and description). However, the time of posting does have an effect on the engagement, which therefore should be taken into consideration.

Data reduction

Based on the data exploration and our own evaluation we have also chosen to reduce the number of variables to the ones we think is relevant for our prediction model and remove the data we won't use. We have chosen the following variables: Tags, date, Numberlikes, NumberFollowers, NumberPosts, NumberFollowing Moreover we have chosen to remove the influencers with most followers to reduce possible errors as they can be seen as celebrities and get a lot of likes on all of their pictures despite the content like the Kardashian sisters etc. Likewise we wanted to lower the number of likes on a picture for the same reason. We chose to limit the maximum followers to 1.000.000, likes to 60000, number of following to 1000 and number of posts to 5000. After this filtering we ended with 10644 observations.

Two models

Base Model

Modeling in machine learning is an iterative phase where we continually will train and test machine learning models to discover the best one for the given task. But before we can begin with the part of testing and training we need to build the model from scratch. When building a prediction model from scratch we will start with building the base model, so we can get a baseline to measure performance against. Our base model is based on integers, which is the variables: Numbers of followers, Number of following and Number of posts. The baseline model showed an accuracy of about 80 % which is an relatively high and good accuracy. Now the question is if the result still will be this high after we include more variables? Since this prediction is based on three out of five variables two credible factors are left out. If we include all the variables we can assume that the prediction model will have a more credible foundation to make a prediction. But this can also lead to a reduction of the accuracy, because more variables can disrupt the coherence between the three..

Deep Learning

We learned in the data exploration that the number of followers and posts not necessarily equals more likes, which therefore indicates the importance of the content of the posts. Therefore we additionally are going to train a deep learning model on tags. The deep learning model therefore is based on the column 'tags' and 'numberLikes', where we want it to predict the tags on posts with likes over the average. It has an average accuracy of about 65 %, which indicates that it is decent at predicting. However, it does need some tuning. In the visualisations of the 10 first predictions, as you can see below, it's clear, that the model is predicting on values that are missing.

7751	[True]	0.00
3669	[False]	nan
1582	[True]	1.00
9861	[False]	nan
8222	[False]	nan
2929	[False]	nan
8712	[False]	nan
6756	[False]	1.00
4883	[False]	1.00
3239	[False]	0.00

We are aware of the fact, that this can have an effect on the accuracy of the model and can be a reason why the accuracy is relatively good. In the assignment we tried to remove the missing values to avoid this possible error, but for some reason the values did not drop and still occurred in the dataset. Therefore it's a possibility that the correct accuracy of the model is lower than 65 %.

CONCLUSION

Finally, we have built a model that is decent at predicting if a post might get above or under 8316 likes based on tags used. However, it does need some more improvement, since there are some issues with missing values in the tag column which occur after preprocessing and may affect the end result.

Additionally, we found out that there is not really a link between number of followers, posts and likes, wherefore the content indicately plays a greater role. This model may then serve as a base for further development, where we recommend to include images and time of posting in the prediction.