











Details	Techniques	Tools	
 Paul Roggeveen  Roggeveen-Analytics  <a href="http://www.roggeveen-analytics.nl">www.roggeveen-analytics.nl</a>  <a href="mailto:info@roggeveen-analytics.nl">info@roggeveen-analytics.nl</a>  <a href="http://www.linkedin.com/in/paulroggeveen">www.linkedin.com/in/paulroggeveen</a>  <a href="https://github.com/Polhovsky">https://github.com/Polhovsky</a>	<ul style="list-style-type: none"> <li>- web-scraping</li> <li>- K-means clustering</li> <li>- Silhouette Analysis</li> <li>- Logistic Regression</li> <li>- Principal Component Analysis (PCA)</li> <li>- Deep Learning</li> <li>- Convolutional Neural Network (CNN)</li> <li>- Transfer learning</li> </ul>	 Python <ul style="list-style-type: none"> <li>- Pandas, Numpy</li> <li>- Matplotlib, Scikit-learn</li> </ul>  Keras  Google Cloud <ul style="list-style-type: none"> <li>- 8 vCPUs, 30 GB</li> <li>- NVIDIA Tesla K80 GPU</li> </ul>	

## SkiNet

Using deep learning to predict the popularity of pictures uploaded by ski resorts

### Abstract

With the enormous popularity of social networks like LinkedIn, Facebook and Instagram, the online world plays a significant role in marketing campaigns. This study focuses on the promotion of ski resorts on Instagram. The official accounts of 80 US ski resorts have been analysed in order to predict the popularity of their pictures with the objective to optimize the use of their Instagram accounts in order to reach the most people.

A state-of-the-art Deep Convolutional Neural Network (DCNN) will be trained to classify the pictures and, together with additional describing features of both the resort and the pictures, will be used for the final prediction. In total over 75 thousand pictures have been used for transfer learning with the VGG architecture in order to optimize the predictions.

A baseline model, without any input from the pictures themselves, achieves an accuracy of 66% while predicting five classes of popularity on a hold-out set. Adding newly engineered features from a DCNN increases the accuracy to 74% for exact predictions and 99% for predictions plus or minus one class.



*Utah's famous powder snow, a dream for many skiers but does such a picture get more likes? (© Snowbird, Utah)*

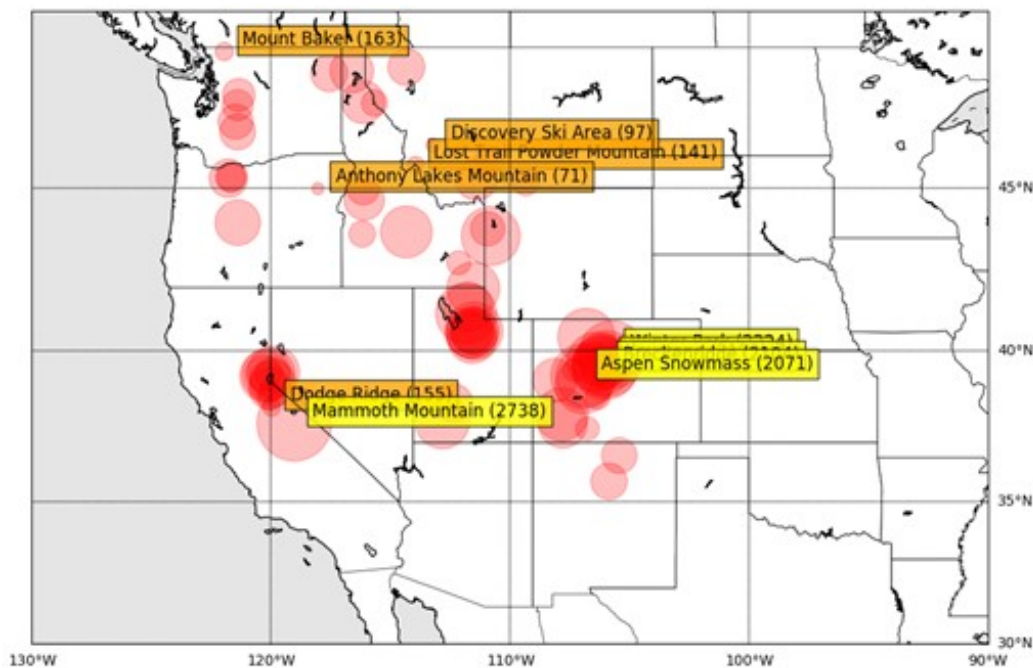
## 1. Introduction

In recent years online social media platforms have exploded both in terms of number of users and activities. As of September 2017 the number of monthly active Instagram users has reached 800 million<sup>1</sup>. Four years earlier this number had only touched 150 million, just to give you an impression of the increase in popularity of Instagram in recent years.

Thousands if not millions of photos are uploaded to these platforms on a daily basis. Consumers use these platforms, for example, to share their favourite restaurants or perfect holiday moments. Research has been done to predict the number of views on Flickr based on a random sample of pictures<sup>2</sup>. However, not only consumers use these platforms but also companies use them for marketing purposes. After all, it's an easy way to reach a large group of people. This study therefore focuses on the prediction of the popularity of images from a business perspective.

This study is performed with data from Instagram, mainly because of the availability of a large amount of pictures. Many companies use Instagram to promote their business but in order to perform a relevant study, a variety of companies is required. An account with over 20 thousand pictures is nice but if the competition of this company isn't very active on Instagram, the results of a study might not generalize very well. In addition, a subject is chosen that speaks to the imagination, or at least mine.

**Figure 1: Number of pictures on the Instagram accounts of Western US ski resorts**



*"Mammoth Mountain and Aspen are amongst the most active resorts on Instagram."*

After exploring several industries and taking into account the requirements of deep learning architectures, I eventually chose to follow my biggest passion: skiing. Training a deep learning model requires a large amount of pictures due to the many parameters or weights of the model and as a result many industries did not qualify for this study. As it turned out ski resorts use their Instagram accounts to promote their business and since there are quite a few resorts, many pictures can be collected. Within the framework of this study the marketing campaign of ski resorts will be analysed and the following question will be answered: what kind of pictures posted by ski resorts will be more likeable and therefore reach a larger audience?

<sup>1</sup> <https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/>

<sup>2</sup> A. Khosla, A. Das Sarma and R. Hamid, What Makes an Image Popular? MIT (2014)

As it turned out many ski resorts worldwide use their Instagram accounts to promote their business. Both Europe and North America have a significant amount of ski resorts that operate throughout the year offering a variety of activities. In order to reduce the cultural influence on the data – people in different regions of the world show different online behavior<sup>3,4</sup> – this study focuses on ski resorts in the United States because of the amount of resorts within the same country posting a significant amount of pictures. This way the cultural influence is limited and at the same time a relatively large amount of pictures can be collected. As can be seen in figure 1, where the most active resorts are posted in yellow and the least active ones in orange, some resorts post more than 2,000 pictures on their Instagram account.

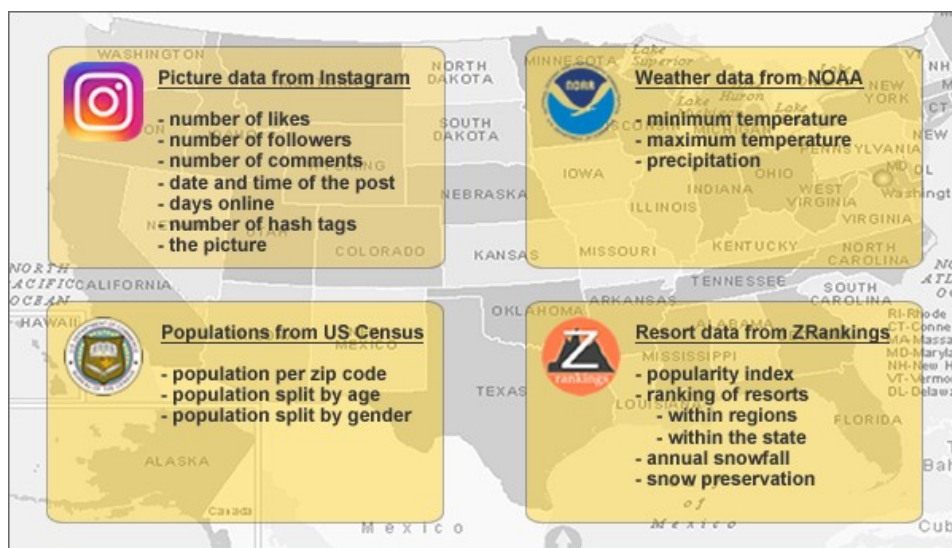
Eventually 75,526 pictures have been collected together with additional data related to both the picture as well as the resort. As a reference a baseline model will be developed to start with. Once insights from this model are gained, the focus will be shifted towards deep learning. With the use of so-called Deep Convolutional Neural Networks (DCNN)<sup>5</sup>, a state-of-the-art technology, the key characteristics of pictures can be extracted in order to classify them. Newly engineered features will be added to the baseline model and is expected to perform best.

In the next sections both data collection and feature engineering, the construction of new features based on the available data, will be explained before discussing the exploration of the data. Once the key characteristics of the data have been pointed out, a strategy to ensure decent generalization of the model will be explained. What follows is an introduction of the baseline model before moving on to deep learning. Finally a summary and the most significant conclusions as well as recommendation for future research will be presented.

## 2. Data collection

Ski resorts exist in all corners of the world, in fact I have visited quite a few of them (visit my personal website [www.polhovsky.com](http://www.polhovsky.com) for more information regarding my ski trips) but I decided to focus on one region in order to minimize cultural differences amongst the Instagram users. In Europe for example social networks are primarily used to share information with friends and family whereas consumers in the United States use social media primarily to express themselves and enhance their image<sup>3</sup>.

**Figure 2: Data sources used to predict the number of likes**



*"Data from several sources are combined in order to strive for the best performance."*

<sup>3</sup> A. Chwialkowska, Is our social media behavior still influenced by our culture? This is how Finns, Poles and Americans differ, University of Vaasa (2017)

<sup>4</sup> L.A. Jackson and J.-L. Wang, Cultural differences in social networking site use: A comparative study of China and the United States, Computers in Human Behavior (2013)

<sup>5</sup> B. Macukow, Neural Networks – State of Art, Brief History, Basic Models and Architecture, IFIP (2016)

As a result this study focuses on the United States only. In total, data from 80 ski resorts, many have posted over a thousand pictures, have been used.

Data collected from Instagram consist of the number of likes, number of followers, number of comments, date and time of the post, days online (days between date of scraping and date of posting), the number of hash tags and, of course, the picture. As you can see from figure 2, the data from Instagram has been enriched with other sources. From the National Oceanic and Atmospheric Administration (NOAA) weather data has been collected which has been used to determine the actual conditions around the date of the post. Data from the US Census has been used to add population data on several levels and split by both gender and age. Finally the popularity of the ski resorts was expected to play a significant role and therefore several KPIs from ZRankings have also been added.

The scraping was completed in the second half of October 2017 and resulted in 75,526 pictures. The objective of this study is to predict the number of likes of photos. The scraped material however contained pictures that were outside the scope of the study. The most frequent amongst these categories are photo collages, watermarked photos of photos with logos, sales ads and weather forecasts. All the photos that were considered to be outside the scope of the study have been manually removed<sup>6</sup>. In total 11,809 pictures have been removed, resulting in a remaining dataset of 63,717 pictures.

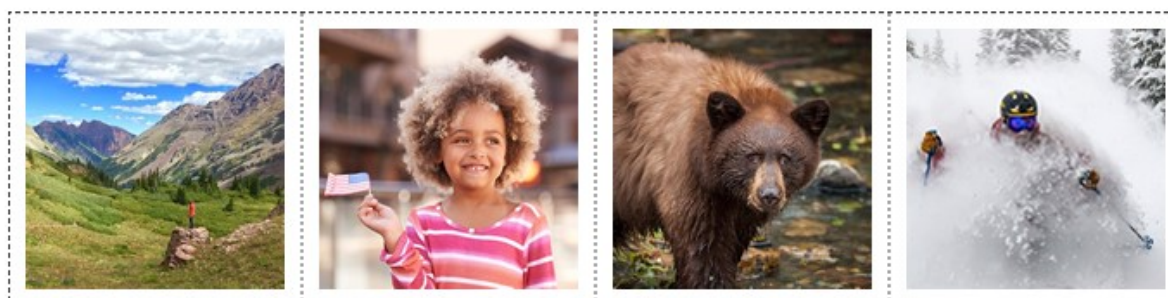
**Figure 3: Examples of pictures that have been excluded from the study**



*"Photo collages, watermarked photos, photos with logos and weather forecasts fall outside the scope of this study."*

The remaining dataset contains photos without added text (signs on mountains or names of restaurants for example do still occur in the dataset). Photos are not modified as some of the examples in figure 3 nor are they screenshots of other photos or radars as is the case with the weather forecasts. What remains are photos of mountains, people, animals and of course, a lot of action shots. Some examples are presented below in figure 4.

**Figure 4: Examples of pictures that have been used for the study**



*"Pictures of mountains, people, animals, and, of course, action shots are included in the dataset."*

Removing pictures manually is, of course, not a flawless procedure. The remaining dataset will therefore contain a few pictures that should have been removed. At the same time some useful

<sup>6</sup> to ensure a low error, the process of manual removal has been checked three times for each resort



pictures might have been removed. However, careful selection has taken place and reviewed several times. As a result the percentage of 'incorrect' pictures is expected to be very low.

The result of the data collection is a dataset containing 63,717 pictures and an abundance of related data from a total of 80 ski resorts in the United States. Cultural differences are expected to be negligible and the remaining pictures are all within the scope of this study.

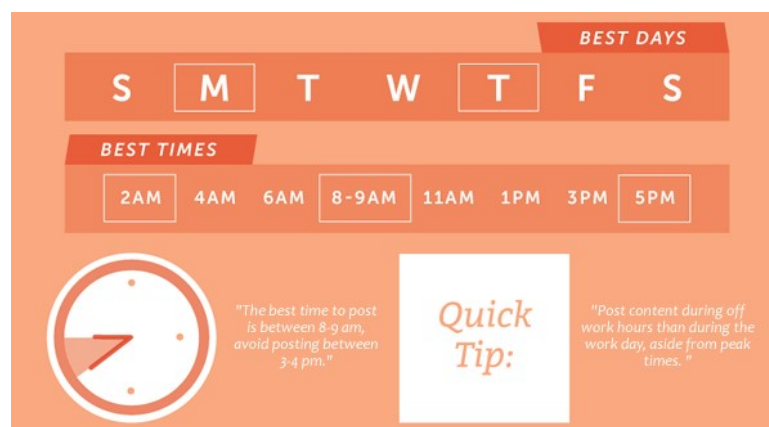
### 3. Feature engineering

As mentioned in the previous section, data from several sources have been collected with the goal to optimize the final predictions. These data sources first have to be transformed into features which can be used as input for the model. This is often referred to as feature engineering. The following features have been developed, or engineered:

- Day of the week, workday, off hours and weekend

The date and the time of posting have been converted from coordinated universal time (UTC) to local time and transformed into several dummies. Day of the week is simply a dummy for each day of the week, for example the feature 'Monday' equals one when the picture has been posted on a Monday etc. The feature 'workday' is a dummy referring to a picture being posted on Monday to Friday between 8am and 6 pm whereas the feature 'off hours' refers to pictures posted on the same weekdays but outside 8am and 6 pm. Weekend is a feature equalling one when a picture has been posted during the weekend. Posting during specific times of the week can lead to more engagement<sup>7</sup>, as can be seen below:

**Figure 5: Posting during specific times of the week relates to engagement**



*"Posting pictures during work hours doesn't only affect productivity, it also won't result in more likes!"*

Figure 5 relates to Instagram in general. For specific branches or industries, like ski resorts, different behavior of Instagram users could be observed. More insights will be presented in the section about data exploration.

- Winter / summer

Ski resorts operate throughout the year with peaks in activity in winter and summer due to holiday seasons and activities. The majority of pictures is uploaded in winter and a small peak in number of posts can be seen in the summer months (figure 8). Two dummies, one for winter and another one for summer have been added. Winter covers the months of November until April whereas summer includes the months of June, July, August and September. May and October are known as shoulder season in the United States with a low number of visitors.

<sup>7</sup> <https://blog.bufferapp.com/instagram-marketing>

- Use of hash tags

Hash tags in a post are expected to have a positive effect on the number of likes. However, using more hash tags will not necessarily lead to more likes. A study performed by Simply Measured<sup>8</sup> showed that posts with at least one hash tag average 12.6% more engagement. Therefore a dummy has been created indicating whether a post contains hash tags or not.

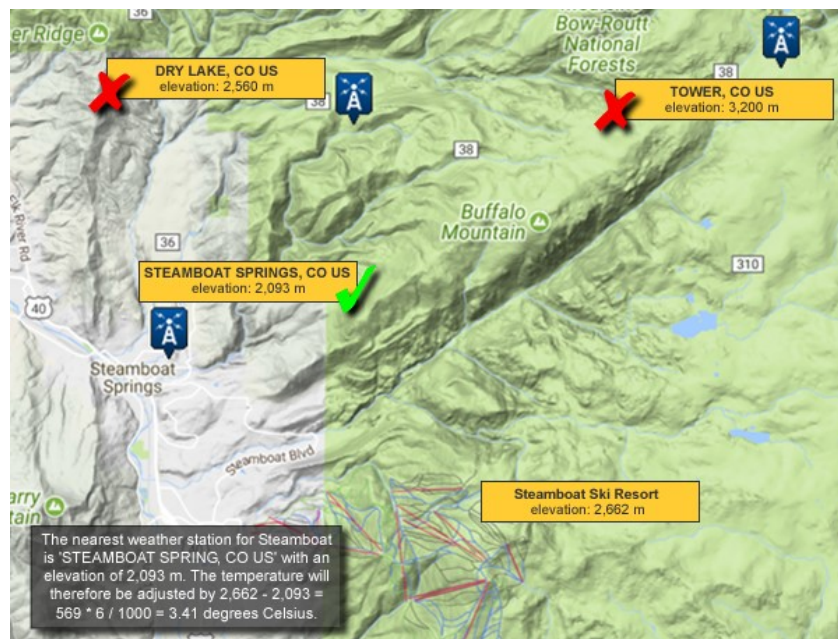
- Population per state

According to digital marketing experts<sup>9</sup>, 59% of the internet users between the age of 18 and 29 use Instagram. In addition the vast majority (68%) of the Instagram users are women. The US Census provides population data of 2010 split by gender and age group for every zip code. After correcting the data for the lag in time<sup>10</sup> from 2010 to 2017 by a factor of nearly 1.05, they have been aggregated resulting in the population of men, women and total between the age of 18 and 29 for every state. A state might not be the exact scope of a ski resort but it serves as a proxy. A larger population in the state is expected to contain more clients of and people interested in a specific ski resort which will lead to more online engagement.

- Minimum and maximum temperature and precipitation around the date of the post

Activity on an Instagram account of a ski resort might show some correlation with weather events. More photos could be posted during a clear day after a storm in the middle of winter when skiing and snowboarding is supposed to be the best. People who experienced such a day might then be interested in checking out the captures of the day.

**Figure 6: Selecting the weather station and correcting for difference in elevation**



*"Steamboat has three weather stations in its vicinity and the nearest one has an elevation which is 569 m lower than Steamboat's mid-mountain."*

From the NOAA three weather KPIs have been collected, minimum - and maximum temperature in degrees Fahrenheit and precipitation in inches. After a conversion to degrees Celsius and mm. respectively, these KPIs have been added to the data in a way that for every picture the weather conditions were available from one day before until one day after the date

<sup>8</sup> Simply Measured, Instagram Study 2014 Q3 (2014)

<sup>9</sup> <https://www.omnicoreagency.com/instagram-statistics/>

<sup>10</sup> <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>

of the post. More weather KPIs would have been highly appreciated, but due to a limited availability from NOAA they could not be added.

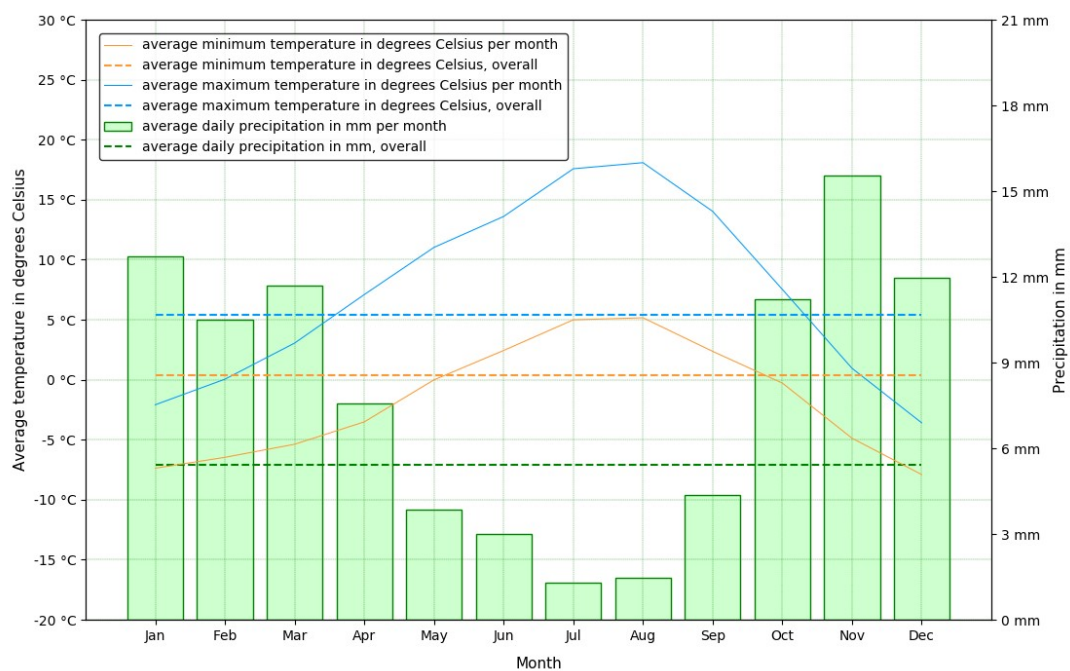
The three KPIs mentioned, were added from the nearest weather station (distances are calculated according to the Haversine formula<sup>11</sup> after, based on the address of the resort, the latitude and longitude had been collected through the Google Maps Geocoding API) with coverage of the specific KPI on the specific date. As a result different KPIs for the same day, or the same KPI for different dates, could be collected from different weather stations. Temperatures were adjusted based on the difference in elevation between the weather station and the mid elevation of the ski resort<sup>12</sup> with 6 degrees Celsius per 1000 meters<sup>13</sup>.

#### - Absolute deviation of the mean weather KPI

As mentioned before, the weather is expected to have an influence on the number of likes of pictures posted by ski resorts. However, care has to be taken here. Ski resorts operate year-round, which means in winter people go out more when it's cold and when there has been recent precipitation. After all, people interested in winter sports enjoy those cold clear days after a storm. In summer however, the opposite might be true. Warm and dry days provide comfort and beautiful views for hikers and bikers. Therefore the absolute deviation from the mean temperature and precipitation has been added to correct for the seasonality in the KPIs.

In the picture below, containing weather data from Stevens Pass, these features are clarified in more detail:

**Figure 7: Weather statistics based on 2011 – 2016 for Stevens Pass**



Minimum and maximum temperatures are low in winter and high in summer, as expected. The absolute deviation from the average will therefore imply that higher values correspond with low temperatures in winter and high temperatures in summer, exactly as required. Precipitation shows a reversed effect, high precipitation in winter and low precipitation in summer. Since more precipitation in winter and low precipitation in summer is usually appreciated by those looking to go out, the absolute deviation is also expected to show a positive correlation with at least the number people going out and possibly with the popularity of pictures related to those days as well.

<sup>11</sup> [https://en.wikipedia.org/wiki/Haversine\\_formula](https://en.wikipedia.org/wiki/Haversine_formula)

<sup>12</sup> <https://www.onthesnow.com/united-states/statistics.html>

<sup>13</sup> [https://www.grc.nasa.gov/www/k-12/problems/Jim\\_Naus/TEMPandALTITUDE\\_ans.htm](https://www.grc.nasa.gov/www/k-12/problems/Jim_Naus/TEMPandALTITUDE_ans.htm),

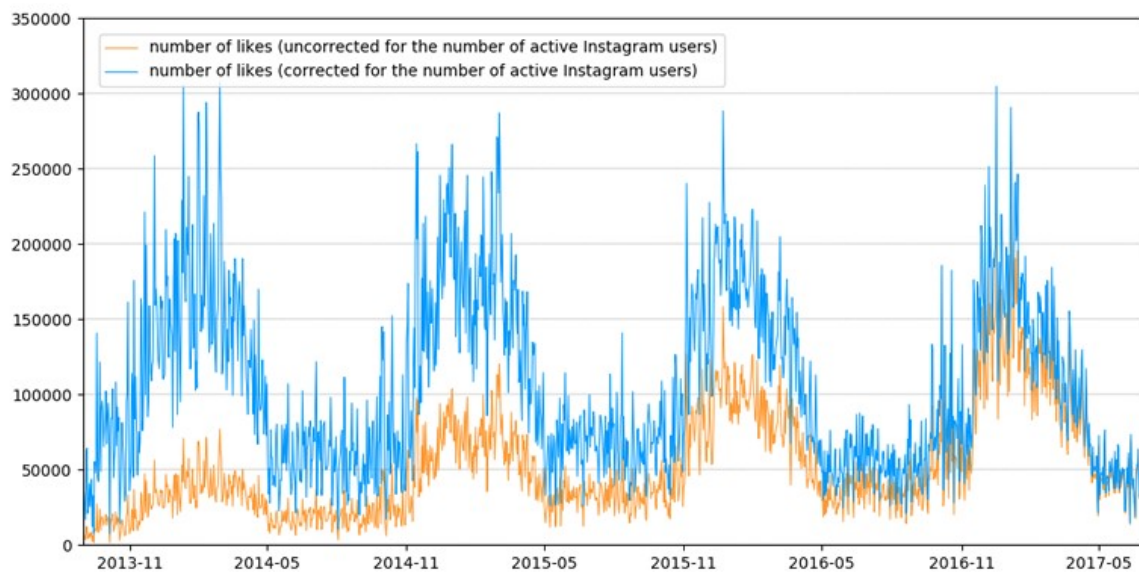
## 4. Data exploration

With all features extracted and engineered, explorations can be done and some visualization can be made. A detailed look at the target variable of this study, the number of likes, is crucial. In addition insights in correlations between the features and the target variable will be established.

First thing to investigate, is the time effect that's clearly present in the data. The number of likes of a picture in June 2014 had a significantly different potential compared to the number of likes of a similar picture posted in the beginning of 2017 due to the number of active users of Instagram. Social media has seen an enormous growth and Instagram is no exception. With 150 million active users worldwide in September 2013 to over 800 million users four years later<sup>1</sup>, this platform is still ranked amongst the most popular social media. Data about the active users have been interpolated and converted to an index under the assumption that the development of users in the United States follows a similar path as can be seen worldwide<sup>14</sup>. This index can now be used to transform the number of likes of any photo on Instagram at any point in time as if it was posted on the same day. Pictures posted before September 2013 have been removed from the analysis due to the fact that the correction factor was too high (Instagram was relatively small at that time) which most likely affects the reliability.

Another thing to consider here is the convergence of the number of likes of a photo to the stabilized, or final, number of likes. A photo receives the majority of its likes in the beginning while after a certain amount of days the number of likes will only sporadically increase. In order to compare the number of likes of pictures they all had to be stabilized or old enough. To be on the safe side, pictures posted after June 2017 (1,834 pictures) have therefore been removed from the analysis. Finally, 27 outliers – pictures without any likes or with more than 1,500 likes (uncorrected for number of active users) have also been excluded from the study resulting in a final dataset containing 52,419 pictures.

**Figure 8: Development of the total number of likes per day for all 80 ski resorts**



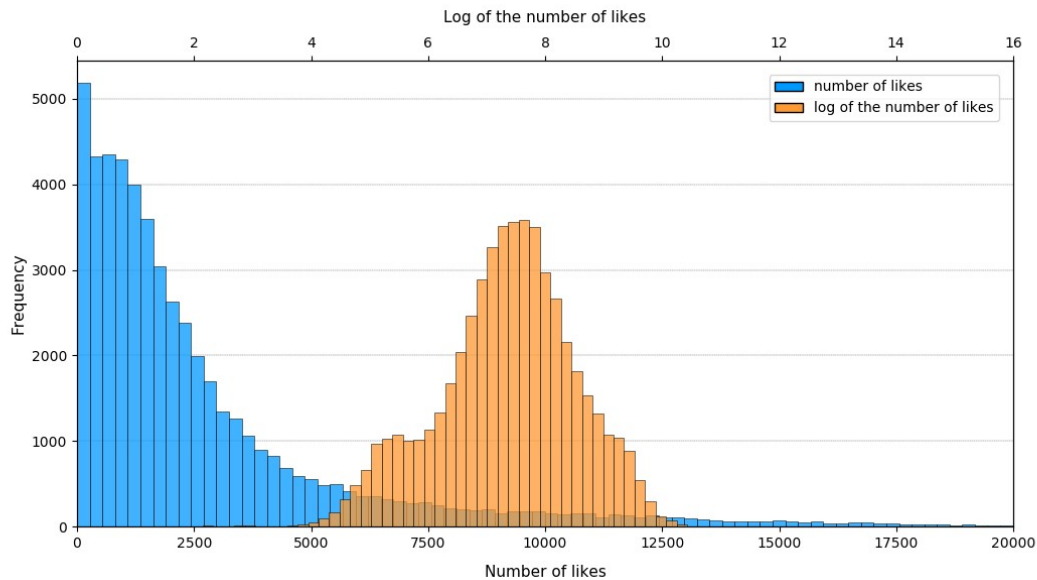
The total number of likes shows an increase over time due to the number of active users (figure 8). However, when corrected for the number of users, the number of likes are constant in time with a strong seasonal pattern. In winter there is clearly more activity than in summer but also in summer a small peak in the number of likes can be seen.

<sup>14</sup> The index equals the interpolated number of active Instagram users divided by the number of active Instagram users in September 2017.



Many pictures do not receive many likes whereas a relatively small number of photos are very popular. This can be seen in figure 9 (blue histogram). The histogram shows a rapid decline in frequency with an increase in the number of likes. The log is a common transformation to convert a distribution such as the one in figure 9 into a normal distribution. The distribution of the log of the number of likes shows a small 'second peak' around 5.5 but nevertheless is close to a normal distribution (orange histogram).

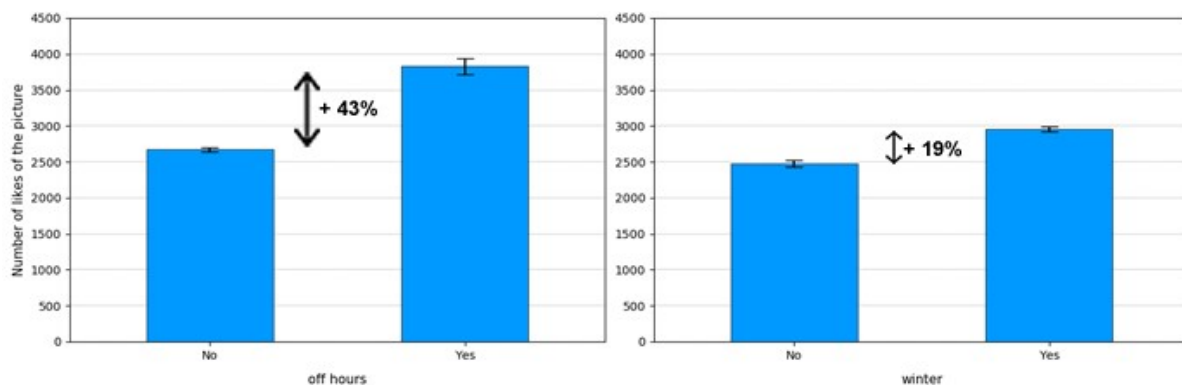
**Figure 9: Histogram of the log of the number of likes**



The next step is to gain some insights in underlying correlations in the data. Since the dataset contains many dummies, correlations are not the best way to indicate dependencies. Therefore the average number of likes between groups has been compared.

The dummies related to the days of the week, working hours, off hours, weekend, winter, summer and use of hash tags have all been analysed. According to figure 5 pictures posted on Monday or Thursday and during off hours will result in more engagement. But, this was based on all pictures posted on Instagram instead of pictures from ski resorts only. This study shows only some effect for Friday till Sunday where posting on Sunday will only result in more likes. Apparently, users interested in ski resorts behave differently regarding online engagement than the average Instagram user. Posting pictures during off hours does result in a significant increase in the number of likes which is in line with the conclusions presented in figure 5. Not surprisingly, pictures posted in the winter also have a significantly higher average number of likes than the ones posted outside winter (figure 10).

**Figure 10: Average number of likes during off hours and in winter**

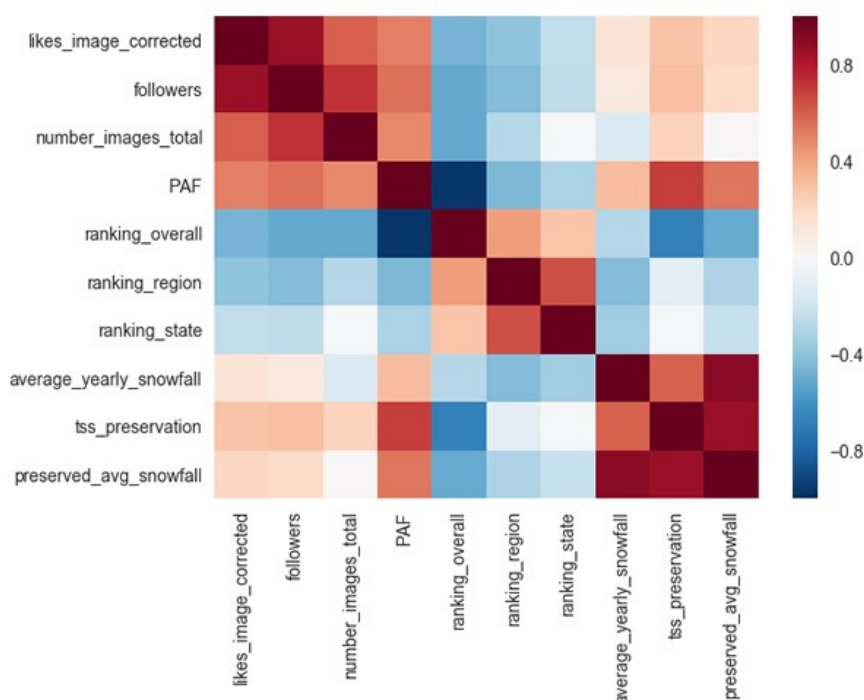


Posting pictures during the off hours will result on average in 43% more likes, a significant increase. Another significant increase can be found when comparing the average number of likes of pictures

posting during winter with pictures posted outside winter. The difference might not be as large as was the case with the previous example, but nevertheless, winter is a dummy expected to add predictive power to the model.

Presenting all the correlations is somewhat comprehensive since there are over 150 features in total. As a result the focus will be on some features that caught the eye. From the correlation matrix below can be seen that the number of followers, total number of pictures posted by a resort and a collection of statistics from ZRankings seem to be correlated, either positively (red) or negatively (blue), with the number of likes. More followers, more pictures posted, a higher popularity index (PAF) and more annual snowfall will lead to more likes whereas a higher ranking number (a less popular resort) has a negative influence on the number of likes a picture will receive.

**Figure 11: Correlation matrix with the number of likes and resort-specific features**



Besides the features discussed in this section, more features will enter the initial model. For example, picture-specific features like the days online (age of the picture) and the hour of the day of the posting will be added. Furthermore, weather statistics and population statistics, as discussed in the previous section, and both dummies for each resort as well as dummies for each state a resort is located in, will be added. Even though the number of comments is heavily correlated with the number of likes, it cannot be added to the model due to data leakage<sup>15</sup>. After all, when posting a picture the number of comments is unknown.

## 5. Generalization

A machine learning algorithm is valuable when it performs well on unseen data. A common methodology is to split the data into a training -, validation - and test set<sup>16</sup> where the first dataset is used to fit the model, the validation set is used to provide an unbiased evaluation of a model fit on the training dataset while tuning the model's hyper parameters and the test set is only used to provide an unbiased evaluation of a final model fit on the training set, indicating the performance. In order to make this work at least the validation and test sets should have similar characteristics. More theoretically, they should come from the same distribution. This is one of the requirements for a machine learning algorithm to generalize well. In this section a method containing histograms and

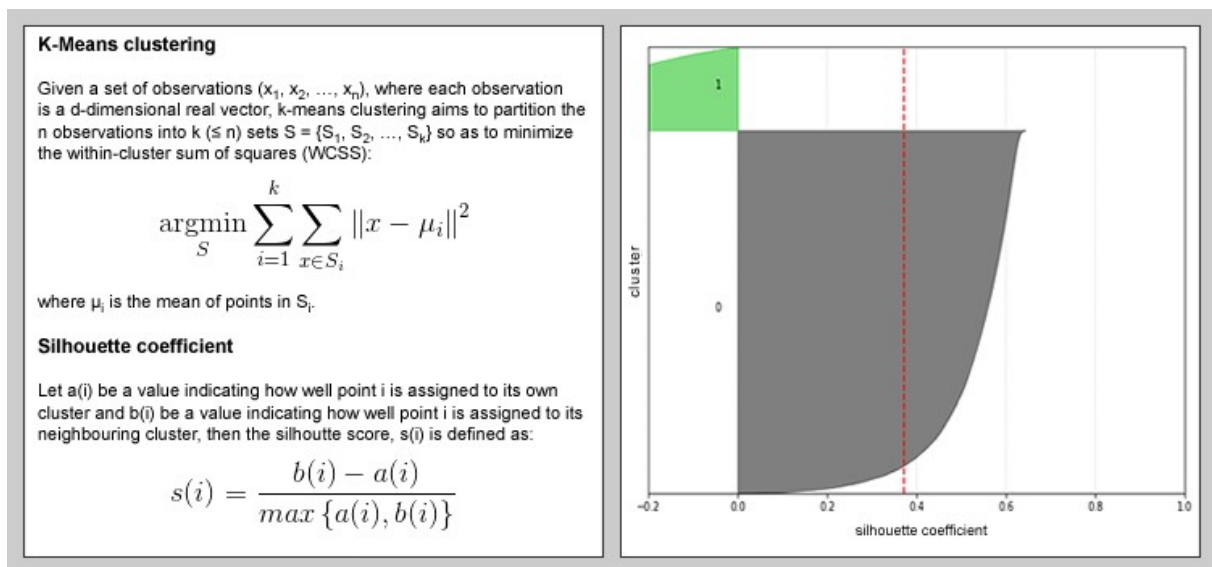
<sup>15</sup> J. Brownlee, Data Leakage in Machine Learning, <https://machinelearningmastery.com> (2016)

<sup>16</sup> J. Brownlee, What is the difference between test and validation datasets?, <https://machinelearningmastery.com> (2017)

cluster analysis will be presented which is used in order to guarantee as much as possible that the model's performance will hold on unseen data.

The pictures extracted from the Instagram accounts of ski resorts consist of several classes of pictures. As mentioned before, the data will be split into three sets of 60%, 20% and 20% respectively. When the different sets would come from the same distribution, the distribution of types of pictures would be comparable in the distinct sets. The first characteristic of a picture that is taken into account for the generalization, is the likability. The corrected number of likes, as discussed in the previous section, has been used to divide the pictures in five groups ranging from the lowest 20% up to the highest 20%. This information has been used to construct the different sets by making sure every set has the same distribution of these five groups. This technique is also known as stratified sampling<sup>17</sup>. Next, the contents of every picture have been used as additional information to further improve the generalization.

**Figure 12: Silhouette plot for K-Means clustering with two clusters**



*"A silhouette coefficient of 0.37 shows the clustering is not optimal."*

In order to classify the pictures into several classes based on its content, a more sophisticated technique like deep learning is required (more about this in section 7). However, with cluster analysis a dataset can be split into several groups or clusters as well. Similar samples, in terms of their distance, are grouped together resulting in a number of clusters. In a good clustering the variance within a cluster is very low whereas the variance between clusters is high. In other words, records in a cluster are very similar and records in different clusters are not similar at all. With K-means clustering one has to input the number of clusters ( $K$ ). Trying several number of clusters, the optimal clustering can be chosen based on the silhouette score<sup>18</sup>.

K-means clustering is computationally expensive and therefore the number of features used as input has to be controlled. Pictures have been cropped (removing borders) and resized to 256 x 256 resulting in a vector of length 65,536 for every channel (colour pictures have three channels: red, green and blue). Next, a histogram<sup>19</sup>, representing the distribution of the intensity of the pixels, has been constructed for every channel. In order to control the number of features, the histograms contain 8 bins where a bin is a range of intensity values. As a result the 3D-histogram of every picture can be converted into a vector of length  $8 \times 8 \times 8 = 512$ .

These 512 features are expected to form at least two clusters. Ski resorts post pictures of scenery and activities both in winter and summer. Since we are looking at the distribution of the colours, a clear

<sup>17</sup> [https://en.wikipedia.org/wiki/Stratified\\_sampling](https://en.wikipedia.org/wiki/Stratified_sampling)

<sup>18</sup> [https://en.wikipedia.org/wiki/Silhouette\\_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))

<sup>19</sup> [https://en.wikipedia.org/wiki/Color\\_histogram](https://en.wikipedia.org/wiki/Color_histogram)

distinction is expected between the histograms of pictures in different seasons. Winter pictures will have a lot of white pixels whereas summer pictures will contain more green.

Even with 512 features K-means turned out to be demanding, so a random subset of 50% of the dataset has been used to detect the clusters. K-means has been carried out for  $K$  is 2, 3 and 4. Based on the silhouette score, which was the highest for  $K = 2$ , the number of clusters has been set. The distribution of colours was clearly not distinctive enough to detect more clusters. The results of the cluster analysis for  $K = 2$  can be seen in figure 12. Since cluster 1 has negative silhouette scores, the variance within this cluster is quite high indicating a variety of pictures, or at least a variety of histograms.

K-means has been fit on the subsample of 50% and has been used to allocate every sample to one of the two clusters. Now let's see if the assumption of a winter and summer cluster is correct. In figure 13 three examples of pictures in cluster 0 are shown:

**Figure 13: Examples of pictures in cluster 0**



*"Cluster 0 can be described as the winter cluster."*

Cluster 0 definitely seems to be a so-called winter cluster. Pictures contain a lot of white pixels or, when looking at the picture on the left, at least do not contain a lot of bright colours. In a similar way, examples of pictures in cluster 1 can be seen in figure 14. These pictures clearly contain more bright colours and could be considered summer pictures or at least non-winter pictures.

**Figure 14: Examples of pictures in cluster 1**



*"Cluster 1 can be described as the summer cluster."*

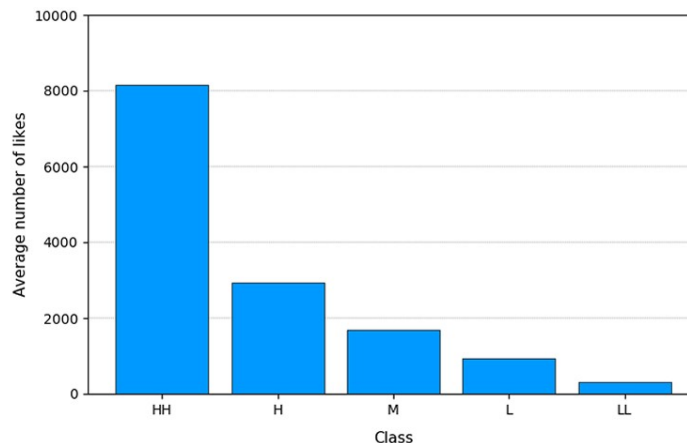
As mentioned before the first step to ensure a good generalization is to make sure the training -, validation - and test set have the same distribution of likability groups. Now, the clusters can also be used to further improve the generalization. Each dataset will also have the same cluster distribution.



## 6. Baseline model

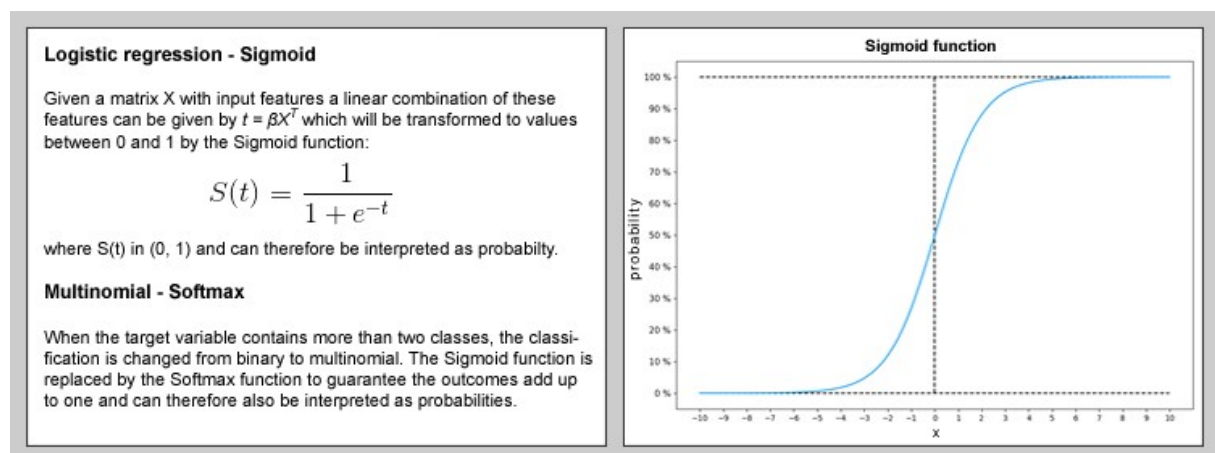
The objective of this study is first to predict the number of likes for pictures posted by ski resorts and second to detect the important factors explaining the popularity of a picture so that ski resorts, or businesses in general, know what to take into account when developing a marketing campaign. As explained previously, the features are related to both the resorts as well as the pictures. The most important features related to the pictures are arguably the pixels composing the picture. These pixels can be used as features in deep learning, which will be discussed in the next section. Deep learning will lead to a higher accuracy of the predictions in many cases but also takes more time to setup and execute. Therefore, first a baseline model will be developed with all the other features as input.

**Figure 15: Average number of likes per target group**



The target variable of the baseline model is the likability, consisting of five groups ranging from the lowest 20% up to the highest 20% of likable pictures (figure 15). The transformation of the continuous variable, the number of likes, into a categorical target feature enables the model to be solved with a classification algorithm. A variety of algorithms is available but due to its interpretability logistic regression<sup>20</sup> has been chosen. Logistic regression can be seen as a regular multivariate regression where the output is transformed into a categorical variable by either a Sigmoid or Softmax function.

**Figure 16: Logistic regression and the Sigmoid - and Softmax function**



The problem at hand is not a binary classification since there are five classes to be predicted. Several solutions exist to tackle this problem. The first one is to estimate five separate models where each one predicts one class versus the other classes. This is also known as 'one-versus-rest' (OVR). A more elegant solution is to predict the different classes at once and that can be achieved by using a

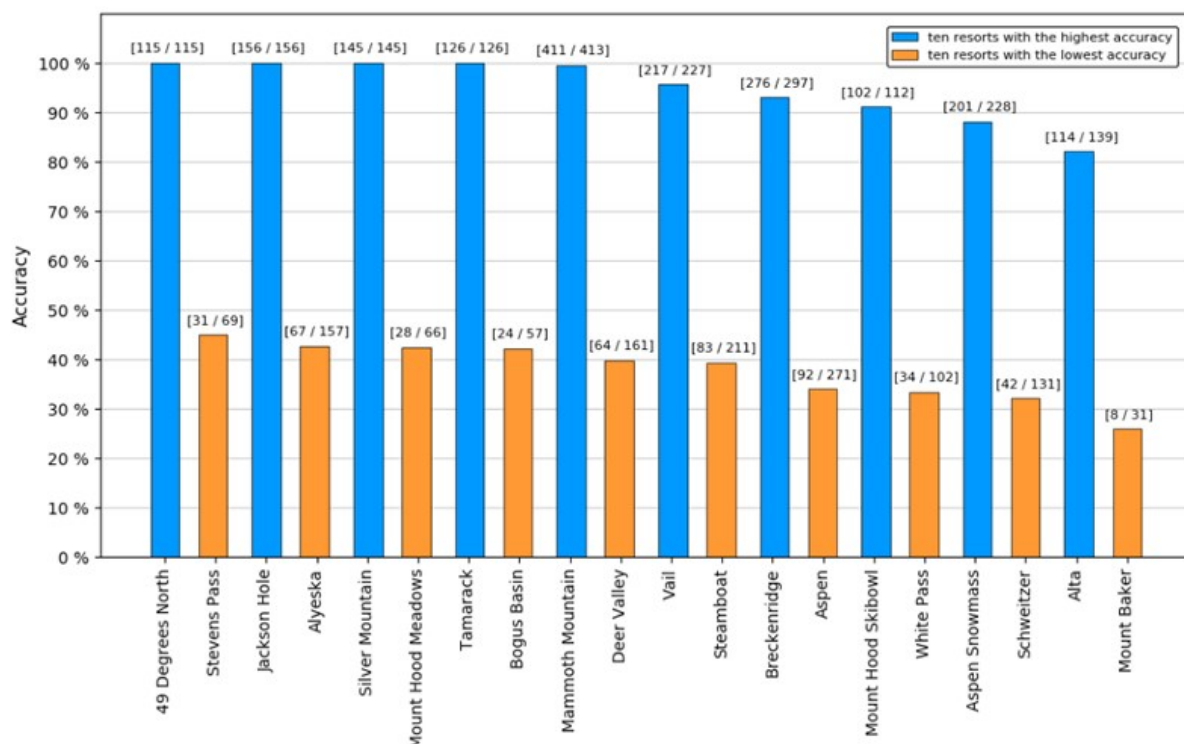
<sup>20</sup> [https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression)

multinomial approach where scores (the output of the multivariate regression associated to each class) are transformed into probabilities by using the Softmax function<sup>21</sup>.

As input for the multinomial logistic regression there are initially 172 features. Before training a logistic regression classifier, a decision tree has been used to reduce the number of features and select the most important ones. Features that don't add any predictive power to the model have been removed, resulting in a selection of 92 features. In contrary to decision trees, logistic regression requires some pre-processing. For the weights of the model to be comparable, standardization is needed<sup>22</sup>. When all features have the same mean and variance and are therefore on the same scale, the size of the weight informs about the importance of a feature.

Some of the remaining 92 features are significantly correlated. Such multicollinearity<sup>23</sup> is undesirable since it affects the reliability of the weights. Once the strongest correlation amongst the features have been removed, 43 features were left as input to the multinomial logistic regression. The algorithm reaches an accuracy of 66% on the training set and 67% and 66% on the validation - and test set respectively. This indicates the model generalizes well since the model is neither over- nor underfitting which happens when a model performs significantly better (or worse) on the training set than on the test set. The most important features are the number of followers, the days online (age of the picture), the popularity index (PAF), the ranking of the resort in the region, total number of images posted by a resort and the hour of the day a picture was posted.

**Figure 17: The ten resorts with the highest / lowest accuracy in the baseline model**  
(number of correct predictions versus number of samples in the test set in brackets)



In order to examine the results of the baseline model a bit more, the accuracy per resort have been plotted. To be more precise, the best and worst ten resorts in terms of accuracy have been plotted in figure 17 where the best ten have been selected from resorts with at least 100 pictures in the test set since a few small (in terms of number of posted pictures) resorts turned out easy to predict because all their pictures were in one class. Notable from figure 17 is the fact that the likability of pictures posted by Aspen Snowmass can be predicted really well with an accuracy of  $201 / 228 = 88\%$  while pictures posted by Aspen have an accuracy of only  $92 / 271 = 34\%$ . Aspen consists of four ski resorts, three of them are represented by Aspen and the other one, Snowmass, is standalone. Since many of the

<sup>21</sup> [https://en.wikipedia.org/wiki/Softmax\\_function](https://en.wikipedia.org/wiki/Softmax_function)

<sup>22</sup> [https://en.wikipedia.org/wiki/Feature\\_scaling](https://en.wikipedia.org/wiki/Feature_scaling)

<sup>23</sup> <https://en.wikipedia.org/wiki/Multicollinearity>

important features are resort-specific, resorts with the majority of pictures in one class are predicted much better than resorts with more variety in terms of the likability of their pictures.

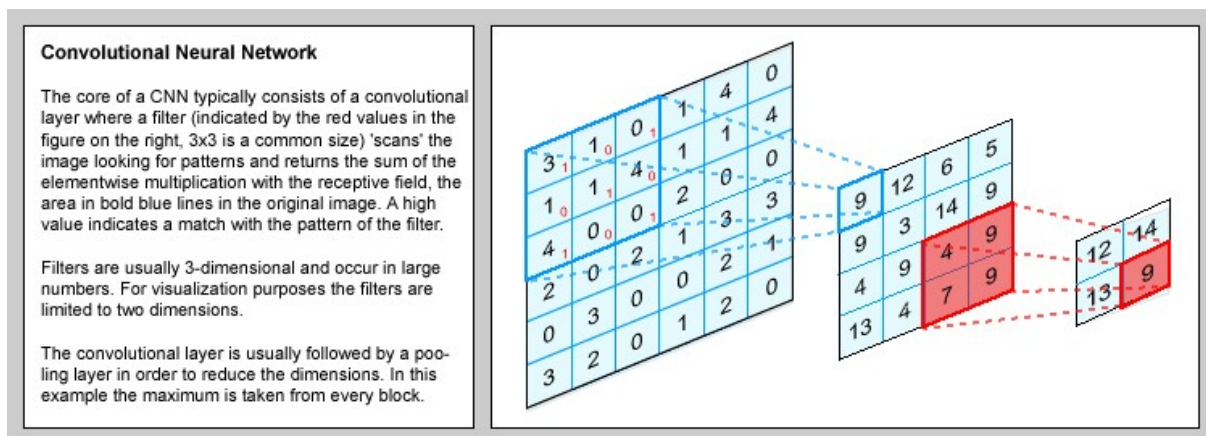
In order to increase the granularity of the model, more picture-specific information has to be added. In the next section the pictures themselves will be used as input for deep learning with the objective to get more insight in the pictures at hand and eventually to increase the accuracy, both overall and on a resort level.

## 7. Deep learning

In this section state-of-the-art deep learning technologies will be explored. While the baseline model was relatively limited to predict with mainly resort-specific features as input, this part of the study is all about adding as much information as possible from the picture itself. First, convolutional neural networks will be introduced before diving into the world of transfer learning. Finally some newly engineered features will be added to the other features in order to train the final model.

Convolutional neural networks (CNN) are a family of neural networks that typically contains several types of layers, one of which is a convolutional layer, as well as a pooling, and an activation layer. The goal of the convolutional layer is filtering, where a kernel moves over an image in order to detect patterns. The pooling layer reduces the dimensions, and therefore the number of weights to train, by extracting the most important information from the convolutional layer. Finally an activation layer makes sure the values are within an acceptable and useful range for the next layer.

**Figure 18: A convolutional neural network: the concepts of filtering and pooling**



One advantage of a CNN is the reduced number of weights that have to be trained due to weight sharing<sup>24</sup>. A stack of filters only contain a fraction of the weights compared to a fully connected layer. And with the objective of image recognition the architecture of CNNs turns out to perform very well. In the early stages of the architecture convolutional layers detect low level features such as curves and edges whereas deeper convolutional layers are capable of detecting high level features such as silhouettes of faces or cars to name a few examples.

The objective here is to engineer extra features in order to improve the accuracy of the baseline model. To be more specific, the pictures posted by the ski resorts have been divided into several classes and the final model will use the class of every picture as additional input for the prediction of the likability of a picture. Image classes were not available, which is where deep learning came in useful. First a random sample of 5,000 pictures had been taken from the 52,419 pictures. This sample will be used to train a model which can then be used to classify every single picture in the full dataset. The pictures in the sample have been manually labeled. After a careful inspection of the random sample, eight classes were defined. More classes are definitely present in the dataset but the size of each class had to be taken into account to guarantee reliable results. Summer and winter activities together with

<sup>24</sup> <http://neuralnetworksanddeeplearning.com/>

summer and winter landscapes are probably some classes that come to mind directly when thinking of ski resorts. Next, people and animals are two other distinct classes of pictures. A remarkable amount of pictures contains lifts making it another distinct class of pictures. Finally, there was a group of pictures left over containing a variety of pictures. They were put together in a category named 'other'. The classes are unbalanced, something that was taken into account during training of the CNN by weighing the classes accordingly<sup>25</sup>.

**Figure 19: The classes of the images**

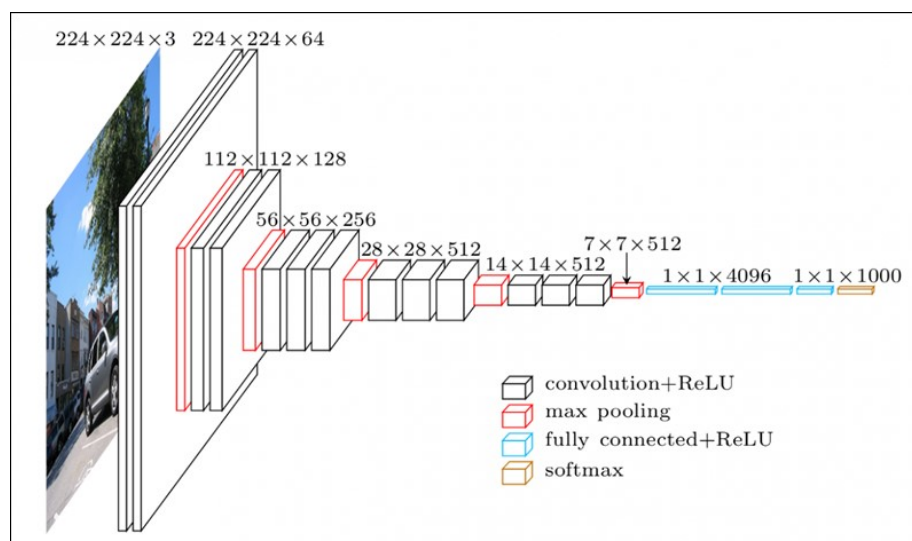


*"From left to right: animals, lifts, other, people, summer activity, summer landscape, winter activity and winter landscape."*

Despite the fact that some pictures could be assigned to several classes, a multi-class model where every picture can only be assigned to one class, was preferred in order to reduce the complexity. In addition it has to be mentioned the labeling will inevitably include some noise. Besides the fact that manually labeling 5,000 pictures almost by nature leads to some errors, some pictures are hard to allocate to one class or the other anyway. Take a scenery in the middle of October after a first snowfall where the mountains are covered in white but the grass at lower elevation is still green, summer or winter landscape?

Before jumping directly into the world of deep learning, a first exploration had been made with so-called principal component analysis (PCA)<sup>26</sup>. The objective of PCA is to reduce the dimensionality by using an orthogonal transformation to convert a set of observations of possibly correlated features into a set of linearly uncorrelated features, the principal components. Colour images of size  $224 \times 224$  were used, resulting in a feature vector of dimension 150,528 ( $224 \times 224 \times 3$ ). By using incremental PCA, a variant of regular PCA capable of handling larger dimensions by using batches instead of the full dataset, 67 principal components can be found representing 80% of the variance of all the features. This enables us to train a regular classifier. Unfortunately the performance of the regular classifier was very poor indicating the principal components didn't contain enough valuable information. As a result either a larger dataset and / or more advanced techniques were required. It turns out both requirements can be met without acquiring additional data.

**Figure 20: VGG16's architecture**



*"VGG16 consists of five blocks of convolutional and pooling layers followed by a few dense layers."*

<sup>25</sup> J. Brownlee, 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset, <https://machinelearningmastery.com> (2015)

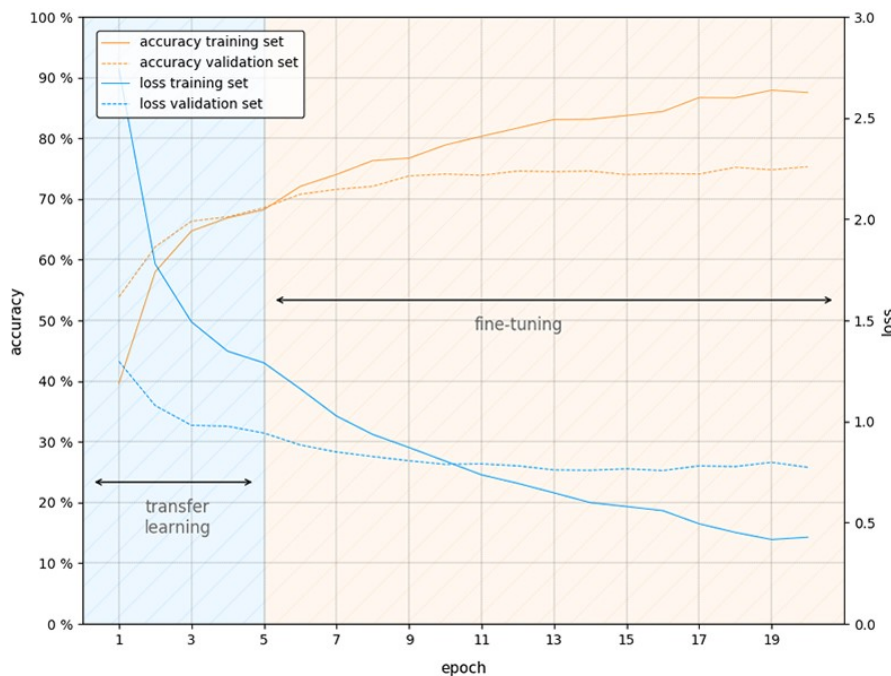
<sup>26</sup> [https://en.wikipedia.org/wiki/Principal\\_component\\_analysis](https://en.wikipedia.org/wiki/Principal_component_analysis)



As mentioned before, a CNN learns to detect certain features depending on the depth of the layer in the network. As it turns out the low level features appear not to be specific to a particular dataset, but in fact they are applicable to many datasets and tasks<sup>27</sup>. Several architectures have been trained on a very large dataset called ImageNet<sup>28</sup>, which are very well suited for transfer learning as this phenomenon is referred to. Every year a contest is being held striving for the highest accuracy of classifying the images from ImageNet into 1,000 classes. Examples of architectures that performed well in these contests are VGG16, Inception and ResNet<sup>29</sup>. Considering that ImageNet contains over 14 million images and the dataset available for this study consists of just over 52K images, it seems only logical to transfer the learnings from a model trained on ImageNet in order to classify the images for this study.

Having tried several architectures, VGG16<sup>30</sup> gave the best results. The architecture of VGG16 is represented in figure 20 and consists of five blocks of convolutional layers followed by a pooling layer. The dataset at hand is relatively small and the content differs from the images in ImageNet so the best strategy is to train only the added customized output layer (which classifies the pictures into eight categories and replaced the original output layer) and, in addition, fine-tune the weights from earlier layers. This way the learnings from VGG16 are used in an optimal way. The model already knows how to detect certain shapes and figures and will learn more from this dataset by fine-tuning the weights.

**Figure 21: Accuracy and loss of the image classifier**



The model has been trained in batches of 32 pictures, all pre-processed as required by VGG16. In addition the training set has been augmented by allowing width - and height shifts of up to 20%, a shear and zoom range of up to 20% and horizontal flipping. Due to its effectivity in recent research, so-called adaptive moment estimation (Adam<sup>31</sup>) has been chosen as the optimizer with a learning rate of 1e-4 during transfer learning (the first 5 epochs as can be seen in figure 21) and 1e-6 during the fine-tune phase (the remaining 15 epochs). An epoch is a complete pass through a dataset. In this case the training set has 3,000 samples, so with a batch size of 32, 93 iterations are needed to complete one epoch.

In figure 21 can be seen that the training - and validation accuracy steadily increase where the validation accuracy stabilizes around 76% in the final epochs. The accuracy of the training set is a bit

<sup>27</sup> J. Yosinski, J. Clune, Y. Bengio and H. Lipson, How transferable are features in deep neural networks?, NIPS (2014)

<sup>28</sup> <http://www.image-net.org/>

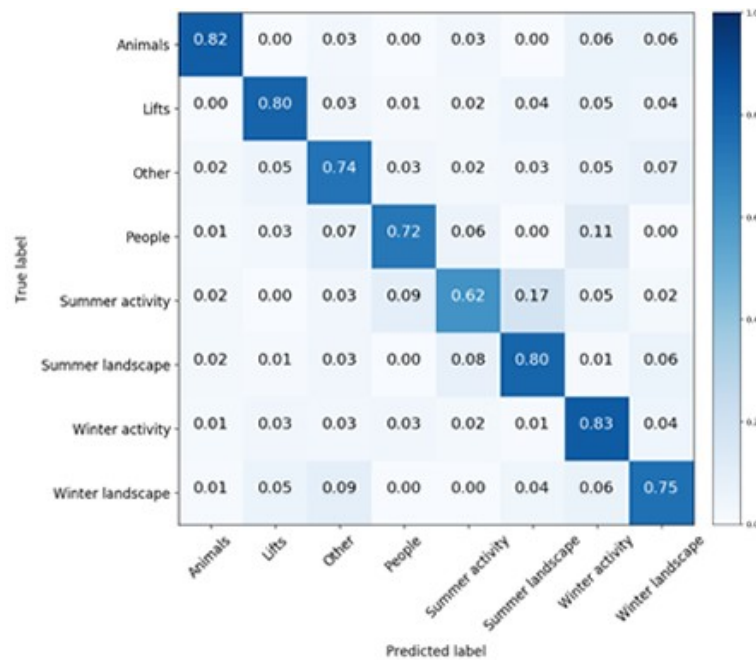
<sup>29</sup> A. Canziani, E. Culurciello and A. Paszke, An analysis of deep neural network models for practical applications, ICLR (2016)

<sup>30</sup> K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR (2014)

<sup>31</sup> D. P. Kingma and J. L. Ba, Adam: A method for stochastic optimization, ICLR (2017)

higher and monotonically increasing which, together with the first observation of a stabilizing accuracy of the validation set, implies some overfitting. Let's have a look at the performance on the test set, data the model had not seen during training:

**Figure 22: Normalized confusion matrix of the image classifier**



The scores on the diagonal represent the percentage of correctly predicted classes. Animals, lifts, summer landscapes and winter activities are all being predicted correctly in at least 80 percent. The model has the most difficulties predicting summer activity. From the samples in this class, 9 and 17 percent are misclassified as people and summer landscape respectively. Another relatively high percentage (11%) of misclassified samples are images labeled as people but classified as winter activity. The overall accuracy of 78% on the test set is a nice result. Let's have a look at both a few correctly and incorrectly classified images:

**Figure 23: Examples of correctly classified images**



**Figure 24: Examples of incorrectly classified images**



In figure 23 two pictures, one labeled as winter activity and the other as people, are shown with their corresponding probabilities. The model predicted the first picture as winter activity with a probability of

76% followed by a 21% probability the picture belongs to the class 'winter landscape'. Given the size of the skier in the photo, one can easily see the confusion. The person on the right is correctly classified in a more convincing way given the probability of 98%.

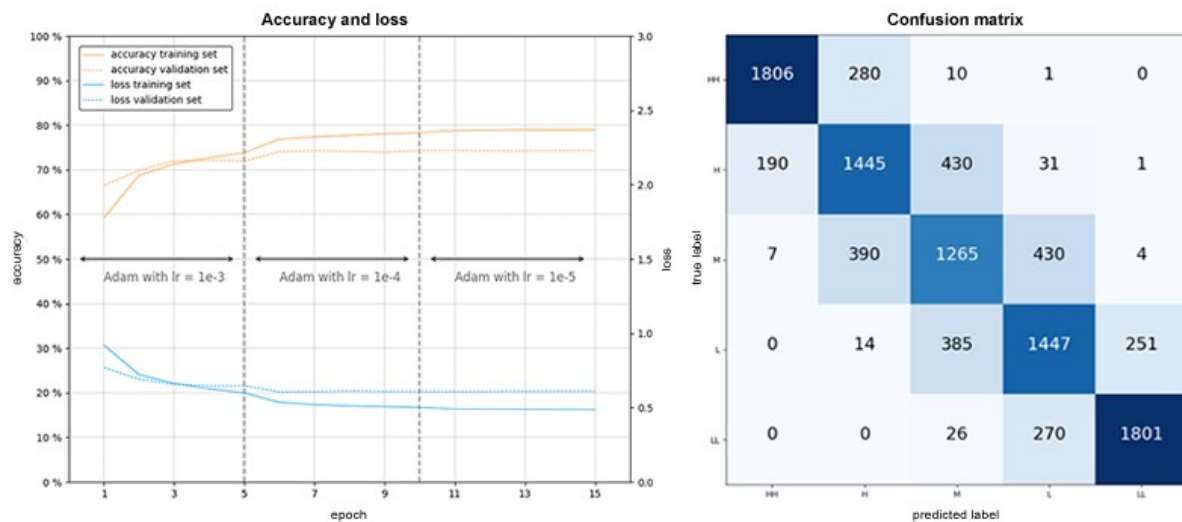
Figure 24 shows two pictures the model classified incorrectly. The picture on the left is labeled as summer activity but the model classified it as winter activity although the second best guess of the model is in fact a summer activity with a probability of 25%. The reason of the highest probability belonging to winter activity might have something to do with the position of the person which is more similar to positions found in winter activities than in summer activities. The picture on the right is even more interesting. A picture of an animal has been classified as winter activity. Perhaps the model has seen many pictures of skiers with wings?

Clearly the model isn't perfect. A larger dataset will be needed to improve the accuracy due to the fact the pictures for this study are so different compared to the content of ImageNet. Nevertheless, thanks to transfer learning a very useful accuracy of 78% can be used to upgrade the baseline model.

The new feature engineered by applying the image classifier on the full dataset has been converted into eight dummies, one for each class, and added to the 43 features that have been used in the baseline model. The accuracy of the predictions increases by only 2% to 68% on the test set. Images of animals, lifts and landscapes (either winter or summer) are likely to receive more likes compared to the other classes of images. Surprisingly, pictures of people and activities are less popular, at least on Instagram. Care has to be taken here with the interpretation of these results since the added dummies only have a marginal effect.

In order to allow for non-linearity<sup>32</sup> (a logistic regression is a generalized linear model), a deep neural network has been trained in order to improve the predictions. After experimenting with several architectures, a network with five hidden layers containing 64 neurons each proved to perform best. This relatively simple architecture trained 20,293 weights which, compared to the 52 weights in a logistic regression, allows for more complexity, besides the already mentioned non-linearity.

**Figure 25: Evaluation of the final model**



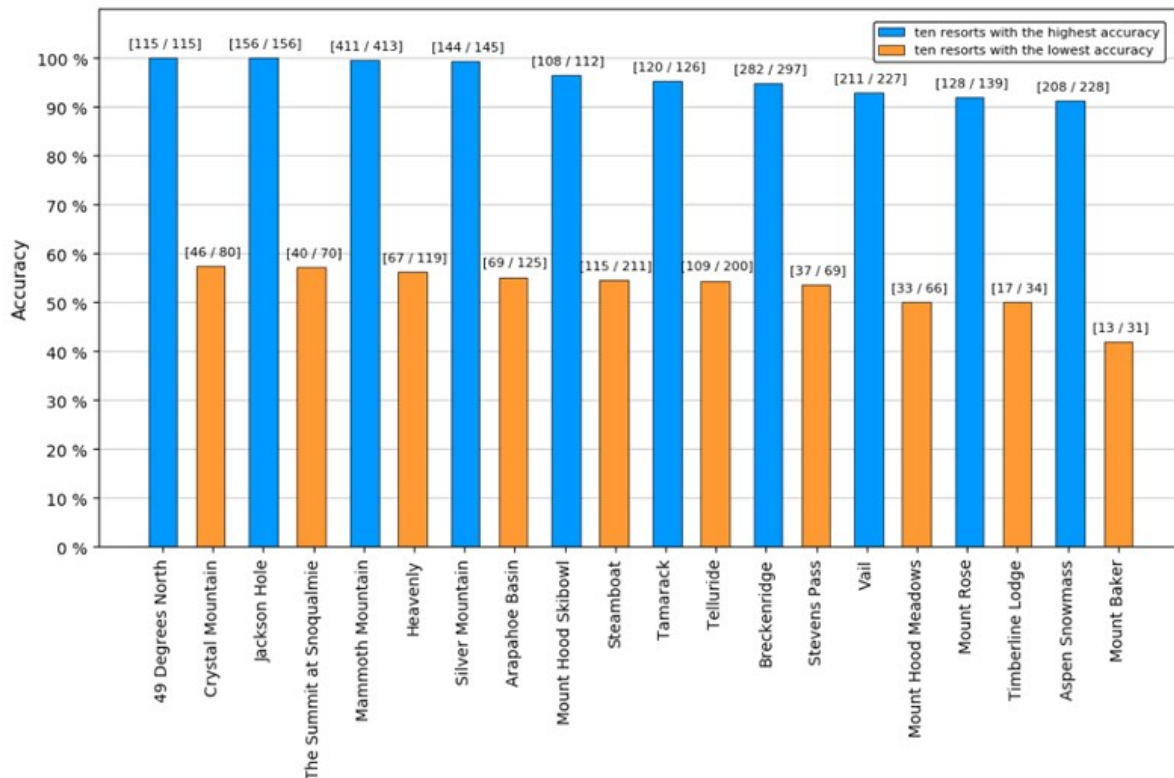
After 15 epochs an accuracy of 79% on the training - and 74% on the validation set was achieved. As you can see in figure 25, an Adam optimizer with a decaying learning rate has been used to train the model. According to the confusion matrix, shown on the right, the accuracy on the test set equals 74%. When allowing the predictions to be less accurate, i.e. they can be one class off, the accuracy goes up to 99%.

<sup>32</sup> S. Dreiseitl and L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, Journal of biomedical informatics (2002)

In figure 17 the accuracy per resort of the baseline model has been analysed. Due to the fact that many of the important features were resort-specific, resorts with the majority of pictures in one class were predicted much better than resorts with more variety in terms of the likability of their pictures. In the final model more picture-specific features have been added and more complex dependencies are allowed. What effect does this have on the accuracy per resort?

**Figure 26: The ten resorts with the highest / lowest accuracy in the final model**

(number of correct predictions versus number of samples in the test set in brackets)



A first observation when comparing figure 26 with figure 17 is the increase of at least ten percentage points in accuracy of the worst predicted resorts. The resort with the lowest accuracy has 13 correctly classified images from a total of 31 pictures (42%) in the final model compared to 8 out of 31 (26%) in the baseline model. When zooming in on the resorts with the highest accuracy one thing to notice is that the overall accuracy went up with all ten resorts scoring above 90% which was about 10 percentage points lower in the baseline model. However, it also has to be noted that the shifts on a more granular level are not all positive. Vail for example went down from 217 to 211 correctly classified images in the final model. Overall, the increase in accuracy from 66% to 74% is reflected on a resort level. The final model significantly improved the predictions on a more granular level but there is still room for improvement. A larger dataset to train the algorithm will probably be most beneficial in this regard.

## 8. Conclusion

In this study a dataset of 52,419 pictures has been composed. Furthermore, data from several external sources have been added in order to enrich the input feature space. A baseline model has been constructed by using a logistic regression in order to classify samples into five categories of increasing likability of pictures. With 43 features an accuracy of 66% was achieved on a dataset the model hadn't seen before. After adding a new feature describing the content of the picture the accuracy increased by two percentage points and allowing more non-linearity and more complexity to the model by training a deep neural network, the accuracy further improved to 74%. When allowing the predictions to be one class off, the final model achieved an accuracy of 99%.



One objective of this study was to predict the likability of pictures posted by ski resorts, another one was to examine which factors drive the popularity of such pictures. The number of followers, the days online (age of the picture), the popularity index (PAF), the ranking of the resort in the region, total number of images posted by a resort and the hour of the day a picture was posted were the most important factors. In addition, pictures of animals, lifts and landscapes showed to be more popular on Instagram compared to pictures of activities although the effects were marginal. So, if you are responsible for online marketing of a ski resort, increasing the number of followers and the popularity of the resort should have more priority than uploading the perfect picture.

In this study several machine learning algorithms have been explored. The data science community being currently very dynamic, new and improved methods will be developed. Some of them might be applicable to this study. However, the biggest improvement can probably be made by obtaining a larger dataset. Thanks to transfer learning the classification of the images was already quite successful but the dataset being significantly different than the one from ImageNet, resulted in an accuracy far from perfect. Increasing the accuracy of the classification from 78% to at least 90% might also increase the significance of the feature in the final model. Other suggestions for future research are multi-input deep learning models combined with transfer learning, ensembling<sup>33</sup> of machine learning algorithms where multiple models are trained and results are averaged and using a multi-crop strategy at test time, so-called multi-crop evaluation<sup>29</sup>, where a classifier is being tested on several different crops of images and again results are averaged.

---

<sup>33</sup> [https://en.wikipedia.org/wiki/Ensemble\\_learning](https://en.wikipedia.org/wiki/Ensemble_learning)