DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING, SECOND CYCLE, 30 CREDITS

STOCKHOLM, SWEDEN 2021

Characterizing Feature Influence and Predicting Video Popularity on YouTube

ALI ABDIHAKIM

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Characterizing Feature Influence and Predicting Video Popularity on YouTube

ALI ABDIHAKIM

Degree Programme in Engineering Physics
Date: July 3, 2021

Supervisor: Somayeh Aghanavesi
Examiner: Aristides Gionis
School of Electrical Engineering and Computer Science
Swedish title: En karakterisering av olika egenskapers inverkan och förutsägelse av videopopularitet på YouTube

Abstract

YouTube is an online video sharing platform where users can distribute and consume video and other types of content. The rapid technological advancement along with the proliferation och technological gadgets has led to the phenomenon of viral videos where videos and content garner hundreds of thousands if not million of views in a short span of time.

This thesis looked at the reason for these viral content, more specifically as it pertains to videos on YouTube. This was done by building a predictor model using two different approaches and extracting important features that causes video popularity. The thesis further observed how the subsequent features impact video popularity via partial dependency plots. The knn model outperformed logistic regression model. The thesis showed, among other things that YouTube channel and title were the most important features followed by comment count, age and video category.

Much research have been done pertaining to popularity prediction, but less on deriving important features and evaluating their impact on popularity. Further research has to be conduced on feature influence, which is paramount to comprehend the causes for content going viral.

Keywords

Viral, popular and trending are interchangeable terms and refers to content that garners hundreds of thousands if not million of views in a short span of time.

Sammanfattning

YouTube är en online-plattform där användare kan distribuera och konsumera video och andra typer av innehåll. Den snabba tekniska utvecklingen tillsammans med spridningen av mobila plattformar har lett till fenomenet virala videor där videor får hundratusentals, om inte miljontals, visningar på kort tid.

I arbetet undersöktes orsaken till virala videor på YouTube. Det gjordes genom att bygga två modeller för att förutspå videopopularitet och därefter analysera viktiga egenskaper som orsakar denna. Resultaten visade att Knn-modellen ger bättre resultat än logistisk regression.

Arbetet visade bland annat att YouTube-kanalen och titeln var de viktigaste egenskaperna som driver popularitet, följt av antal kommentarer på en video, videons ålder och videons kategori. Vidare forskning är dock nödvändig inom detta område. Mycket forskning har gjorts för att förutsäga populariteten hos videor, men mindre fokus har lagts på att analysera deras viktiga egenskaper och utvärdera deras inverkan på populariteten.

Nyckelord

Viralt, populärt och trendigt är utbytbara termer och syftar på videor som får hundratusentals om inte miljontals visningar på kort tid.

Acknowledgments

I would like to thank my supervisor, Somayeh Aghanavesi, whose expertise helped in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Stockholm, June 2021
Ali Abdihakim

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

YouTube is a video sharing platform that provides content for users to consume, mainly in the forms of videos which are published from content providers. The available content includes video clips, TV show clips, music videos, short and documentary films, audio recordings, movie trailers, live streams, video blogs, and short original videos. Most content is generated by individuals, but media corporations also publish videos. Besides watching and uploading content, registered users can comment on videos, like and dislike them, create playlists, and subscribe to other users. YouTube is also the second most viewed website in the world as of 2020 [2].

YouTube's rapid growth as a video consumption medium has proven to be a highly effective tool for content providers to share and distribute their content while video distributing platform such as Facebook, Instagram and TikTok have certainly aided YouTube's rapid growth and development. This rapid advancement, along with the proliferation of technological gadgets such as mobiles, smartphones, pads and computers has led to a phenomenon called viral content. This phenomenon describes how a piece of content garner views and audience in an unprecedented manner, often garnering hundreds of thousand if not millions of views in days.

The potential impact of such viral videos can therefore be immense and their impact on multiple aspects of society are undeniable. This can be noticed in content marketing for instance, which is a staple for many businesses to attract and garner attention from their customers in order to turn the viewership into potential profit, hence the reason businesses target their customers via short YouTube ads. Viral videos have also had a profound social impact

on other aspects of society, such as politics. For instance, during the 2008 US presidential election, the pro-Obama video “Yes we can” went viral and received approximately 10 million views [3]. This is not to trivialize the US presidential election to a viral piece of content, but to show the financial, marketing and societal impact of viral videos. The impact of viral content has therefore attracted the attention of researchers and predicting content popularity has become an essential focus in academia.

It is therefore paramount to take a closer and more analytical look on what constitutes such a piece of content. This is the aim of this study. The goal of this thesis is to understands the features that cause videos on YouTube to become popular. It is also to observe how those features influence predictive outcome. To do this is a predictor model was built using two methods, logistic regression and k-nearest neighbour (knn). The two methods performance were compared and feature importance was extracted using the models. Further evaluation was done to not only observe feature importance, but to measure how those features impact the outcome, meaning the relationship between features and target value, namely popularity.

1.2 Problem

1.2.1 Problem Statement

From an engineering perspective, building a popularity predictor model in and of itself is not highly complicated. However for the model to be proficient in real life application one has to consider the multiple exterior components that affect the popularity of a video. Trends heavily impact whether a particular type of video will become widespread or not and a trend could be anything from a song to a dance move to a clothing line. Considering these influences can complicate the predictor model. In addition, features importance would fluctuate based on current events and whats trending at the time.

While there are other research regarding popularity predictor models, not as much has been done to understand the causes for videos going viral, especially pertaining to video features and their importance to driving video popularity. Even less research has been conducted to further analyze the relationship between video features and predictive outcome. If YouTube content providers could predict whether their video will be trending or not and properly gauge feature importance and their respective relationship with driving video popularity, they could adjust their videos to garner the most attention from the public. This is currently difficult to do since YouTube have

not been transparent with their method of evaluating trending content.

1.2.2 Research Question

The research questions in this study are:

- What features in a YouTube video are most important to drive popularity.?
- What are the relationships between those features and video popularity?

1.3 Goals

The goal of exploring the research questions has been divided to three practical sub-goals:

- Build a predictor model with two methods, logistic regression and knn and compare their performance.
- Extract feature importance to video popularity.
- Evaluate relationship between features and popularity.

1.4 Purpose

There are two main purposes of this thesis. The first one is to build a popularity predictor model with two methods, logistic regression and knn and investigate and compare their performance with suitable evaluation methods. The second purpose is to use the built model to extract and examine feature importance to video popularity. The third purpose is to measure how those features influence the models outcome. This project will therefore be beneficial to content marketers, for further sociological studies and for researchers in my field, namely computer science who have and will continue researching connections between video popularity and machine learning modelling.

1.5 Social and Ethical Factors

The notion of predicting popularity and feature importance can be leveraged in multiple ways and can thus have tremendous ethical implications and societal. Assuming that a proficient enough predictor model is available for public use,

or even for governmental bodies, such a predictor could be used for profit and other business purposes. However, a large ethical implication is how it can be utilized to propagate misinformation and disinformation. Media outlets had previously been nicknamed the fourth state because of their ability to frame and propagate political issues despite not being formally recognized as a political system. A model like this could be used to reinforce that notion. If used by the public, then viral videos with disinformation would most likely circulate all over the web. If it is constrained for state and governmental systems, then one can suspect a large scale of strategically circulated viral video to reinforce certain narratives for what can be considered mass brainwashing.

1.6 Delimitations

The delimitations for this project are many since there are tens of research question that can be derived from the project and there are also multiple components of a YouTube video one can dissect and analyze further. For instance, an entire research question can revolve around the title of a video and how the structure of the title, and choice of word could affect a videos popularity. One could do this for every single data point such as thumbnail, description, views etc. The scope of the degree project would be far to big and could hamper its quality with such an approach. There are also numerous ways of building a predictor model. One could build a model that aims to predict number of views, while another aims to predict whether a video will go viral or not, focusing on a more classification approach. The point is that the limitations are placed in order make the project focused and reachable.

In this project the aim is to build a predictor model using a classification approach. This strictly limits the scope of the project and puts focus the main research questions.

1.7 Structure of the thesis

chapter 2 presents relevant background and related work. In chapter 3 the method is described and results are presented in chapter 4. The discussion is placed in chapter 5 and chapter 6 concludes the thesis.

Chapter 2

Background

This chapter provides background information about the methods used for building the model logistic regression and knn. This chapter also discusses the evaluation method used called ROC-curve which is a graphical plot which illustrates the performance of classification methods. Feature importance is also discussed in more detail. The chapter also describes related work.

2.1 Logistic regression

2.1.1 Linear Regression

It is important to grasp the concepts of linear regression before transitioning to the theory and understanding och logistic regression. Logistic regression is after all an extension of the linear regression model suited for classification problems.

linear regression is a famous and well known modelling approach that aims to predict an outcome based on the weighted some of the feature inputs. It can also be used to model a linear dependence between targets y and features x [4].

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{2.1}$$

Observing equation 2.1, the outcome y is a weighted sum of its x features. The betas are the coefficients, often time learned weights with a particular learning rule. There are multiple ways to determining the optimal beta. The least square

methods is a well known and often used rule. Epsilon is the error parameters. It highlights the difference between the ideal outcome and the actual outcome.

The linear regressions distinct property is in its linearity. It forces the model to be linear as seen with equation 2.1. This can be described as its greatest weakness and strength. Linear models are easy to understand, mathematically straightforward and easy to implement since there are an abundance of resources of software, libraries and implementations that one can utilize.

The issue is that the model can only be represented linearly, hence it is excessively dependent on the distribution of data one is going to predict. This is why linear models generally have poor predictive capabilities. The model often oversimplifies what in reality is much more complex [5].

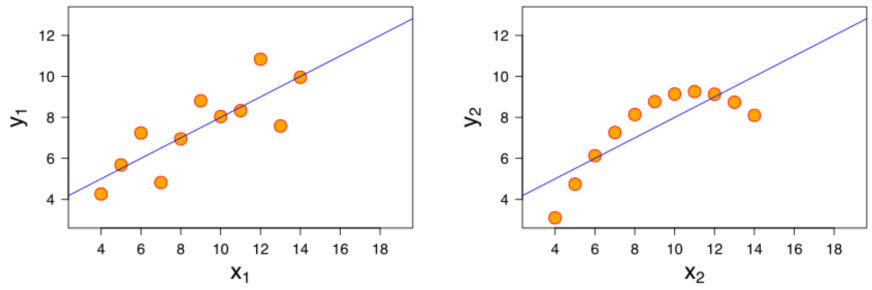


Figure 2.1: Example of linear regression. Figure to left shows appropriate modelling. Figure to right shows poor modelling, linear regression is not appropriate. All figures in thesis are from Wikipedia and are copyright free with no attribution required.

While a lot more can be said on the topic, this short introduction to linear regression is sufficient to understand logistic regression, which is paramount for this thesis.

2.1.2 Logistic Regression

While linear regression works well for linear models, it fails for classification problems. This is because it does not output probabilities. Imagine having classes 0 and 1. A linear model will construct the optimal hyper plane that minimizes the euclidean distance (see section 2.2.2 for more on euclidean distance and distance metricises). However with a classification instance we want to develop a predictive model with a classes 0 and 1 and a probability distribution regarding the classes and the outcome.

This is where logistic regression comes in. Logistic regression is a mathematical statistical method by which one can analyze measurement data. The method is best suited for investigating whether there is a relationship between a response variable (Y), which can only assume two possible values, and an explanatory variable (X) [6]. For instance:

If one is interested in observing whether there is a relationship between the amount of tar in the lungs (X) and whether one has lung cancer (Y). The response variable can only assume the two values, 'Yes' or 'No', while the explanatory variable can assume any positive values.

This is in juxtaposition to the well known linear regression that we discussed which is an approach to modelling the relationship between a scalar response and one or more explanatory variables.

With logistic regression, we are interested in a relationship between the probability that Y will assume the value 'Yes', and the explanatory variable X

$$Prob(Y = Yes) = f(X) \tag{2.2}$$

The logistic regression model uses the logistic function for classification purposes and squeezes the output of a linear equation between 0 and 1. The logistic function is defined as:

$$logistic(x) = \frac{1}{1 + exp(x)} \tag{2.3}$$

and it looks like this:

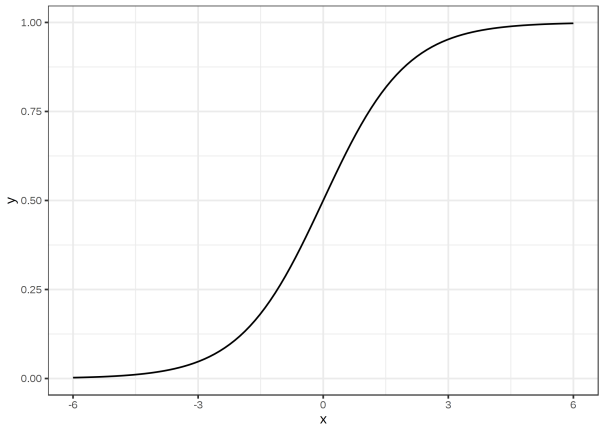


Figure 2.2: The logistic function. The output is between 0 and 1, and output is 0.5 at 0.

The process of going from linear regression (eq 2.1) to logistic regression

is intuitive. While the linear regression models the relationship between outcome and features linearly, the logistic regression outcomes ranges between probabilities 0 and 1. To do this, we wrap the right side of eq 2.1 into the logistic function (eq 2.3). This causes the values to only range between 0 and 1 [7].

$$logistic(x) = \frac{P(y=1)}{1 + exp(B_0 + B_1x_1 + B_2x_2 + ... + B_nx_n)} \tag{2.4}$$

2.2 K-Nearest Neighbor

K-nearest neighbor is a well known unsupervised machine learning method used for both regression and classification. It is a non-parametric learning algorithm, meaning that there is no presumption regarding the data distribution and the data does not necessarily need to have a well-known distribution. Knn is also known as a lazy learning algorithms or a just-in time algorithm. This implies that generalization occurs after a query is made as oppose to when the algorithm generalizes the data before receiving queries, which is the case for eager learning algorithms [8].

Knn is a good classification method and despite its simplicity in both theory and applicability yields highly favourable results. It is also as previously mentioned a non-parametric algorithm, thus not requiring assumptions regrading data distribution which is notably convenient when the data is unusual in its distribution. One also has the option of choosing the distance criteria, depending on which distance metric one chooses to implement. While the euclidean distance is the most commonly used distance function, there are also alternative methods such as hamming distance that calculate the distance between binary vectors, a method more suited for categorical features [8].

However, there are a few issues with the knn algorithm. One of such being that it does not have a training phase, thus requiring every classification of new data to search for its nearest neighbour in the complete training set. The algorithm's efficiency also rapidly decreases as the dataset grows. Another pertinent issue is regarding missing data points. Although there are a few tricks to manage the lack of data, real world data often has a great deal of missing data.

2.2.1 Knn Classification

The concept in knn is to classify an object based on the multiplicity of votes of its neighbouring objects. It closely resembles the "apple does not fall far from the tree" idiom, noting that a child often times has similar traits or characterises to their parents because of their close proximity. Similarly, the algorithm classes a new data point based on how its neighbouring data are classified. The algorithm does this by looking at k number of already classified data points and classifies the new data point based on majority votes from its k-nearest neighbours [9].

Since Knn algorithms is based on a majority vote from k number of classified data points It is paramount to:

- Correctly choose the value of parameter k.
- Choose an appropriate distance metric to calculate distance from new data point to classified data points (often euclidean distance).

K is a very important parameter in knn and it is therefore a vital problem when implementing the method. The practice of correctly selecting the value of k is called parameter tuning and can have a substantial effect in classification accuracy. If k is chosen to be small, then there is the risk for overfitting, causing poor generalization and accuracy on unseen data. However, if k is to larger then there is the probability that the model will require high computational resources. Figure 2.3 shows a good illustration of knn and how value k can impact the classification.

While there are no predefined methods for finding the optimal value of k, selecting $\sqrt{n} = k$ where n where n is the total amount of data points works relatively well [10]. However, selecting the value of k depends on individual cases. A good practice is therefore to run through multiple values of k and compare and verify the outcome. Cross-validation can be used for this, and picking the k value that results in good accuracy can be considered the optimal k value for the particular case.

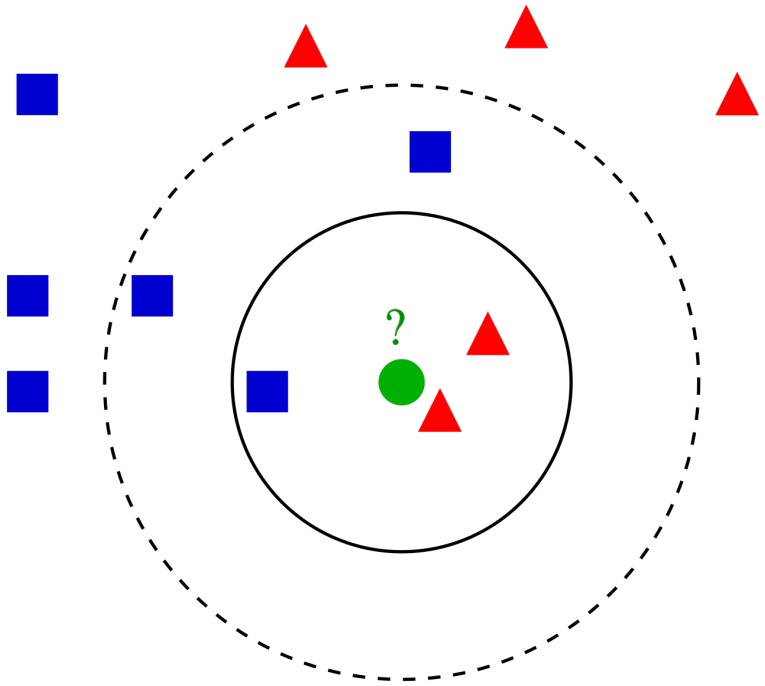


Figure 2.3: Figure illustrates a knn classification example [1]. The green dot is the test sample that is to be classified to either a blue square or a red triangle. If k = 3 (solid inner circle) then the green dot is assigned to the red triangle since there are 2 red triangles versus 1 blue square inside the inner circle. However is k = 5 (dashed outer circle), then the green dot is assigned to the blue squares since there are 3 blue squares versus 2 red triangles inside the outer dashed circle.

2.2.2 Distance metric

The distance metric, the function to calculate the distance to fellow neighbouring data goes hand in hand with knn. The most notable of such distance metric is the euclidean distance with the following formula.

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2} \quad (2.5)$$

The idea is to calculate the distance between the unclassified data and all the classified data points resulting in an array of the distance between the unclassified data to all the other data already classified [11].

2.3 Evaluation Methods

There are many evaluation methods for measuring the performance for classification models, however ROC-AUC curve was chosen along with model accuracy as evaluation methods for predictor performance. ROC-AUC is suited in situation with an uneven and skewed sample distribution. Since in our case the non-trending data is smaller then the trending-data used (more on this in section 3), ROC-AUC is a pertinent evaluation metric.

2.3.1 ROC-AUC

ROC-AUC stands for Receiver Operating Characteristic and Area Under The Curve and is an evaluation measurement used for measuring classification models performance. The ROC-AUC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. TPR stands for True Positive Rate and is calculated with:

$$TPR = \frac{TP}{TP + FN} \quad (2.6)$$

FPR stands for False Positive Rate and is calculated with:

$$FPR = \frac{FP}{FP + TN} \quad (2.7)$$

- TP = true positives: positive values predicted as positive

- FP = false positives: negative values predicted as positive
- FN = false negatives: positive values predicted as negative
- TN = true negative: negative values predicted as negative

A ROC curve is a probability curve an AUC is the area under that curve. AUC is a value between 0 and 1 and represents the a separability measure and indicates the models predictive capabilities, the higher the AUC, the better the models correctly classify data [12].

2.4 Feature Importance

Extracting feature importance aims to assign each feature a score based on their usefulness at predicting a target value and is an important part in predictive modelling. It helps provide insight into the data and the model and can aid in further improving the predictive model. There are multiple feature importance methods, in this thesis permutation feature importance was employed.

While feature importance measures which features mostly affect the predictions, partial dependence plots evaluates how certain features affects predictions. A partial dependence plot shows the relationship between features and the respective target. It works by marginalizing the models output over all features, other than the features that are of interest and computes the relationship between the features of interest and the predictive outcome.

2.4.1 Permutation Feature Importance

The concept is intuitive and easy to follow. Permutation feature importance works by permuting the feature's values and grade feature importance based on the increased prediction error. A feature is considered more important the larger the prediction error become. Thus, if scuffling a feature's value leaves the error unchanged, then that feature is considered unimportant.

There are many ways to shuffle the data. Some suggest a methodical approach of permuting each feature value i with every other feature value j. Fisher et al. recommended to divide the data in half and swap feature values [13]. Then there is the issue of computing feature importance on the test or training data. Not much research has been done on this. One needs to decide

whether to focus on the models ability to measure each features importance on unseen, test data or to prioritize analyzing each features importance to the models predictive performance using training data.

2.4.2 Partial Dependency

Let x_s be the features considered, x_c be the other non-interesting feature, f be the machine learning model and n be the number of data in the data set. Note that x_s and x_c constitutes the whole feature space. The idea is to marginalize over the non-interesting feature x_c thus getting a function dependent on only x_s features. The partial dependency function is estimated with:

$$\hat{f}_{x_s}(x_s) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_s, x_c^i) \tag{2.8}$$

It is common practice to only consider one or two interesting features at a time and observe the relationship and affect on video popularity. The partial dependence plot displays the marginal impact of the feature to the target for multiple values of the feature. E.g. if the line is at 0, then for that value of the feature variable, there is 0 impact to the target.

2.5 Related work

There is a growing amount of research carried out to predict the popularity of distributed content over social media due to their prevalence in society. Below are a sample of studies performed to predict and analyze content popularity.

2.5.1 YouTube Videos Prediction: Will this video be popular?

In 2019, the Department of Civil and Environmental Engineering, Stanford University researched YouTube video popularity prediction [14]. The videos were divided into four classes: non-popular, overwhelming praise, overwhelming bad views, and neutral videos. The selected features for the algorithm

were title, time gap, category, tags, description, and duration. They also implemented multiple machine learning algorithms to compare predictive performance:

- Stochastic gradient descent (SGD).
- Multi-layer perceptron (MLP).
- Decision Tree and Random Forest
- Gradient Boosting Method and Extreme Gradient Boosting

They also observed feature impact on video popularity. They found that extreme gradient boosting performed best with an f1-score of 0.736 and features category and description had the most impact on video popularity.

2.5.2 Predicting popularity of online videos using Support Vector Regression

In May 2017, the following paper aimed to propose a regression method for video popularity, namely support vector machine (SVM) with Gaussian radial basis functions, that would challenge the state-of-the-art regression implementations such as, Univariate Linear (UL) Regression, Multivariate Linear (ML) Regression and Multivariate Radial Basis Function (MRBF) Regression. The purposed method was evaluated on three datasets, combining of approximately 24,000 videos, and the results showed its marginal superiority over the state of the art . [15].

Moreover, the paper evaluated and compared the influence of social features (views, comments, etc) and visual features (thumbnail, title, etc) on video popularity and showed that the social features represent a better signal in terms of video popularity prediction than the visual ones.

2.5.3 Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study

The following paper was authored in 2017 and the authors aimed to present a model to predict videos with logistic regression while adopting a stepwise regression method to improve its accuracy [16]. The logistic regression with the stepwise regression had an improved accuracy from 91.10 to 91.82 %. The

paper also presented a popularity metric that incorporates multiple parameters that indicate video popularity such as views, comments etc. The metrics was based on the following formula [16]:

$$Pr_{pop} = \alpha * mean(views) + \beta * mean(comments) + \gamma * mean(rating) + \sigma * mean(rate) \tag{2.9}$$

The coefficients are weighted in relation to feature importance. For instance, alpha will be a larger coefficient than beta since the number of views is a more important parameter than the number of comment. The popularity metric was computed and if a video had a higher popularity metric than the one computed with function 2.9, then it is considered popular. If not, then it is not considered popular.

2.5.4 Predicting the popularity of online content

This paper presented a method for predicting future popularity performance of YouTube videos and Digg news stories based on early performance [17]. The purpose was to demonstrate that future content popularity can be predicted shortly after submission by measuring popularity at an early time. The results showed that measuring performance of Digg stories two hours after submission allowed the authors to forecast the stories popularity 30 days ahead with good accuracy. 10 days of following measurements of a YouTube video was needed for the same accuracy. The authors mention the differences in content type as a reason for different time scales needed for good prediction. They argue that YouTube videos are not as time contingent as Digg stories since news quickly become outdated while YouTube videos can easily be found long after initial submission.

Chapter 3

Research Methodology & Implementation

3.1 Data Collection

The dataset for trending videos was collected from Kaggle, an online community of data science and machine learning practitioners with resources for computer scientist to use for their projects. Note that Kaggle datasets often comes with projects and solutions related to the data, I therefore had to formulate the thesis to adjust to this fact such that there is no preexisting solution to the research questions, and nothing similar to that extent. The trending dataset included several months of trending videos on YouTube and was updated daily with approximately 200 videos added daily from India, USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea, and, Japan respectively. The videos were uploaded during year 2018, from 1 January 2018 to 31 December 2018. The data included YouTube video features, meaning data attributed to a single YouTube video. Those features were:

- **video age** - Age of the video. This was calculated using the "publish time" feature which says when the video was published.
- **trending date** - Date in which the video was first trending.
- **title** - Video title.
- **channel** - Name of the channel that published the video.
- **description** - Description of the video.

- **category** - Video category associated with the video.
- **tags** - Tags associated with video.
- **views** - Number of views on the video
- **likes** - Number of viewers that liked the video.
- **dislikes** - Number of viewers that disliked the video.
- **comment count** - Number of comments on the video.
- **trending** - Value that tells if video is trending or not. Values equals to 1 if video is popular/trending and 0 if it is non-trending.

The dataset was in csv format and contained approximately 25.000 videos. The missing values were filled with zeros.

While there already were existing trending YouTube video dataset, there were no preexisting data for non-trending videos, something needed to train the models with. YouTube uses a combination of factors and user interaction measurements (views, likes, comments etc) to determine whether a video is trending or not. By "non-trending" videos, I simply mean videos from 1 January 2018 to 31 December 2018, the same span as the trending videos, that were not trending during that year, meaning they were not in the trending dataset. The non-trending dataset was collected using the YouTube API, allowing users easy access to scrap video metadata. One selects video ID and queries the YouTube API. Also, while queering, one receives multiple more features than the above-mentioned ones such as privacy status, licensed content and many more.

3.1.1 Pre-processing

When queering for the non-trending dataset one receives multiple additional features such as video privacy status, etc. The non-trending data was therefore reformatted to be consistent with the trending data set. This was done by manual deletion of features that were not present in the trending dataset. The non-trending data contained 8400 videos. The trending data and non-trending data were then combined into one large csv file to be used for the models.

Notice that all data points were numerical values with the exception of the channel, title, tags, description and category features. Dummy variables were used to handle the category feature which is a categorical feature. The category IDs were mapped to their corresponding category names using the JSON

dictionary file that came with the data. Separate columns were then created for every category and if a video belonged to a category it was represented by 1, otherwise it was represented by 0. There were 15 video categories, see table 3.1 below for them and their respective IDs. That leaves features channel, title, tags and description that needs to be represented numerically. This was done with tf-idf, term frequency–inverse document frequency, which is a method for computing the importance of words to a document. The idea is that the importance of a word increases proportionally to the number of times that word appears in the data, but the importance is also offset by the frequency of the word in the entire dataset. We will come to this, but the data needed to be processed beforehand.

All upper case letters were normalized to lower case letters. YouTube channels, titles, descriptions and tags often consists of numerous common word in all videos. These are called stop words and are words such as to, the, if, not, etc and they were therefore filtered and removed and unique words were considered. The text was then split into separate words and tokenized, meaning that the repetition of words were counted. If a title had a word repeated twice, that would give that specific word a value of 2, this process was repeated for every word and is called tokenization. Tokenization represents the number of times a term appears in a feature data (video title, tags, description). This value was used to compute tf-idf.

Now back to tf-idf. Computing tf-idf was done by computing the tf, the idf, and then multiplying tf*idf and adding all the values of each word for each feature. For instance, if a video description consisted of 5 words and each word had a tf-idf value of 3, that would give that description a value of 3*5=15 which would replace the description text. This was done for the four features channel, title, tags and description for all videos. The formulas for computing tf and idf are below.

$$Tf(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ feature\ data}{Total\ number\ of\ terms\ in\ the\ feature\ data} \tag{3.1}$$

$$Idf(t) = \ln\left(\frac{Total\ number\ of\ data\ (videos)}{Number\ of\ data\ (videos)\ with\ term\ t\ in\ it}\right) \tag{3.2}$$

All data had been represented in numerical/categorical form and was used to train and evaluate the model. The data was a 33391 x 27 size csv file. 33391 is the number of videos, trending and non-trending data combined. 27 is the number of features which are the 12 above mentioned features and 15 dummy

columns for the categorical feature, category (see table 3.1).

Categories is an especially essential feature that was closely observe. This is since feature importance for videos of each category were observed. Table 3.1 below presents available categories attributed to a YouTube video.

ID	Category
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	How-to & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism

Table 3.1: Table shows YouTube video category IDs and their respective categories.

3.2 Implementation and Evaluation

The implementation of both k-nn and logistic regression algorithms were written in Python. The sci-kit library was used during the whole implementation, from preprocessing to modelling. A large part of the implementation was regarding the data and managing and processing the data too use for the model which was discussed in more depth above, in section 3.1.1. Additional libraries such as numpy, pandas and matplotlib was also used for data management and visualizing the results. Logistic regression was implemented with a 80-20 train/test split. A knn classifier was also implemented and compared to the logistic regression method. The models were then evaluated with ROC curve and their respective accuracy was also computed.

Other than building a predictor model, extracting and analyzing important features for video popularity was also part of the research question. Permutation

feature importance is a technique for calculating relative feature importance and can be used independently of the model. The model was fit on the dataset and the model made a prediction on the data and repeated it for each feature multiple times. The result was the mean importance score for each input feature, hence feature importance. This was done to extract the feature importance for the whole dataset. This was also implemented for data of each video category separately. Table 4.3 in the result section presents video feature importance over individual categories.

Additionally, the goal was to not only observe what features were important but how they were important. Partial dependence plot were used for this purpose. The logistic model was fit, and only one feature of interest at a time was considered initially. thereafter two features at a time were considered.

Chapter 4

Results

In this section the results are presented. The accuracy and score of the models logistic regression and knn are presented. Section 4.2 present feature importance to video popularity and 4.3 displays partial dependence plots.

4.1 Model Performance

Table 4.1 and 4.2 presents performance of the algorithms in terms of accuracy. Table 4.1 considers the accuracy for the knn-model for multiple values of k and the table below presents the average knn accuracy along with the accuracy of the logistic regression model. Accuracy follows the formula below:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalSample} \tag{4.1}$$

K-value	Accuracy (%)
5	91.17
10	90.63
15	90.55
25	90.53
50	91.03
100	90.32
200	89.97
500	87.95

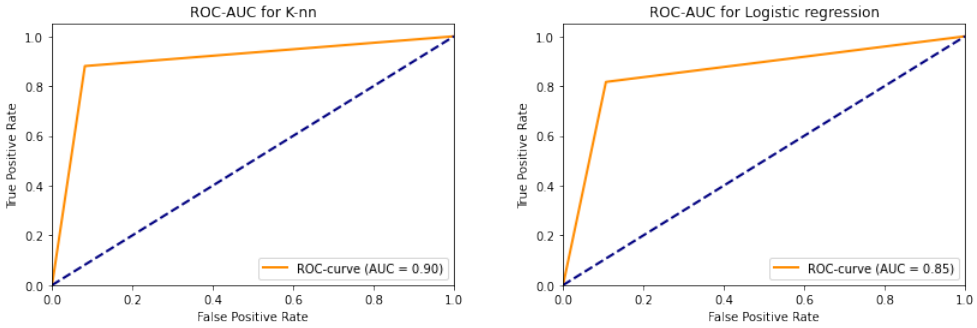
Table 4.1: Presents the accuracy of knn with multiple values of k.

Model	Accuracy (%)
Average knn performance	90.27
Logistic regression	84.39

Table 4.2: Accuracy of knn (average) and logistic regression.

Knn had an average accuracy of 0.903 and logistic regression had an accuracy of 0.844. Knn also performed slight worse with k > 100. In fact the highest accuracy was with k = 5, the smallest k, while the lowest accuracy was with k = 500, the highest k value.

Figure 4.1 a) and b) displays the ROC-curve for knn and logistic regression implementation respectively.



a) K-nearest neighbour. b) Logistic regression.

Figure 4.1: Presents ROC-AUC for both knn and logistic regression models. AUC displays area under the curve.

Knn also performed better than logistic regression, according to ROC-AUC curve, AUC was 0.9 for k-nn and 0.85 for logistic regression.

4.2 Feature Importance

This section presents a diagram with features and their relevance to video popularity. Some features importance were negligible and are not shown in the figure below. The feature importance figure estimated channel as the most important features, followed subsequently by features title, age, comment count and category. Views was the second least important feature.

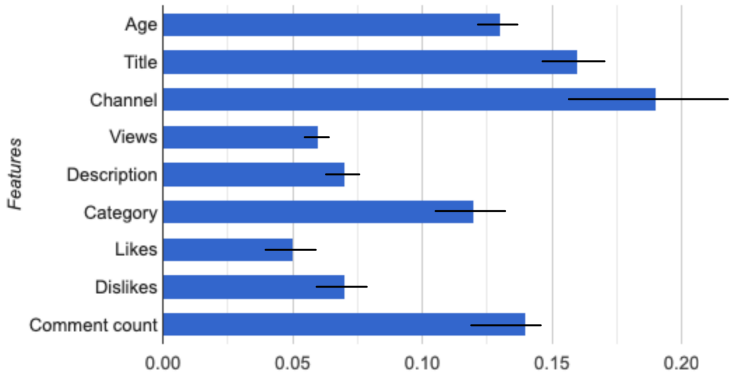


Figure 4.2: Diagram shows features and their importance to video popularity with their respective error bar.

4.2.1 Feature Importance Over Categories

This section presents a table with feature importance for all YouTube video categories. Results that differ considerable with figure 4.2 are highlighted with bold text.

Categories	Feature Importance								
	Age	Title	Channel	Views	Description	Category	Likes	Dislikes	Comment count
Film & Animation	0.15	0.16	0.19	0.06	0.05	0.12	0.05	0.07	0.14
Autos & Vehicles	0.13	0.16	0.19	0.07	0.06	0.12	0.05	0.07	0.14
Music	0.19	0.20	0.31	0.06	0.0	0.05	0.02	0.07	0.09
Pets & Animals	0.12	0.24	0.21	0.06	0.07	0.12	0.05	0.07	0.05
Sports	0.17	0.26	0.11	0.06	0.13	0.08	0.02	0.04	0.12
Travel & Events	0.13	0.16	0.19	0.06	0.07	0.12	0.05	0.07	0.14
Gaming	0.10	0.13	0.25	0.06	0.07	0.12	0.05	0.07	0.14
People & Blogs	0.13	0.16	0.19	0.06	0.07	0.12	0.05	0.07	0.14
Comedy	0.13	0.16	0.18	0.07	0.07	0.12	0.05	0.07	0.14
Entertainment	0.14	0.16	0.19	0.05	0.07	0.12	0.05	0.07	0.14
News & Politics	0.25	0.20	0.20	0.06	0.09	0.10	0.01	0.01	0.07
How-to & Style	0.13	0.16	0.19	0.06	0.07	0.12	0.05	0.07	0.14
Education	0.13	0.16	0.19	0.06	0.07	0.12	0.05	0.07	0.14
Science & Technology	0.13	0.16	0.19	0.06	0.07	0.12	0.05	0.07	0.14
Nonprofits & Activism	0.13	0.16	0.19	0.06	0.07	0.12	0.05	0.07	0.14

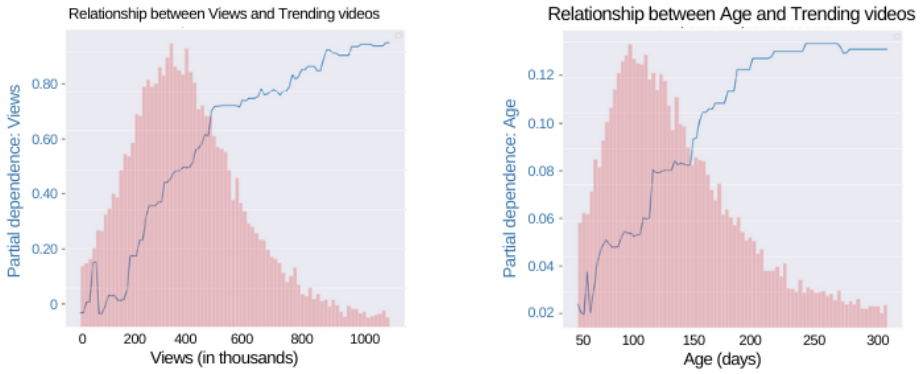
Table 4.3: Feature importance over all video categories.

Table 4.3 shows that of the 15 categories, 4 of them had a stark feature importance difference to the figure 4.3. The categories were music, pets & animals, sport and news & politics. The differences can be summarized as follows:

- All categories put more emphasise on video title.
- Music, pets animals and news politics categories were more influenced by the channel feature while sports category deemphasized the importance of the channel.
- Sports category put more concern on video description and music put less concern on the same feature.

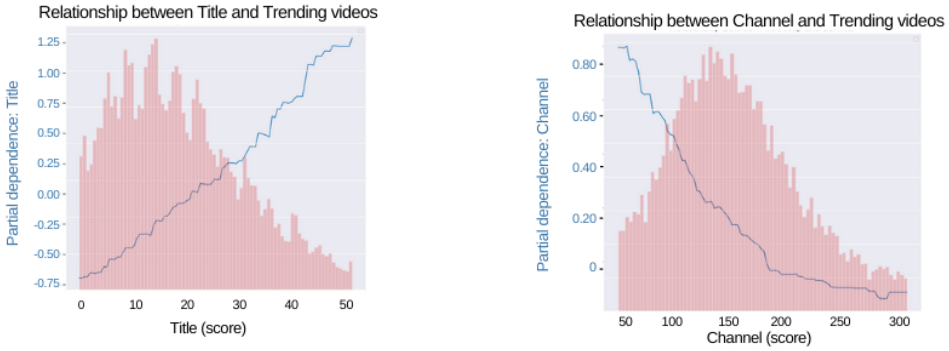
4.3 Partial Dependence

The partial dependence plots are presented in this section. Features considered are age, title, channel and views, excluding features category, description, likes, dislikes and comment count. This is done since likes, dislikes and comment count are strongly correlated to the number of views while and video description is correlated to the title. Category is a categorical feature and the distribution and impact were even through all categories. It is therefore sufficient to observe partial dependence on features age, title, channel and views while excluding the others, plotting them then becomes tedious and redundant.



a) Views. b) Age.

Figure 4.3: Figures displays partial dependency of plot for features a) views and b) age. The red histograms shows the distribution



a) Title. b) Channel.

Figure 4.4: Figures displays partial dependency of plot for features a) title and b) channel. The red histograms shows the distribution

Observing figure 4.3 a) and b) one can see an initial linear relationship between feature value and feature impact, however that linear growth starts to taper of approximately midway through the x-axis. Figure 4.3 a) distribution shows that there are few videos with over 800 thousand views and b) shows that few videos are older than 250 days. The impact of the view feature grows as the amount of views becomes larger. Interestingly the same can be said for the video age. The impact of the age feature grows as the videos ages.

Figure 4.4a) displays a linear relationship between title score and impact on video popularity. Figure 4.4b) on the other hand shows an almost inverse linear relationship. The higher channel score, the less marginal affect it has on video popularity. The distributions in both instances are somewhat similar, most videos are in the middle scores (10-40 for title, 100-250 for channel).

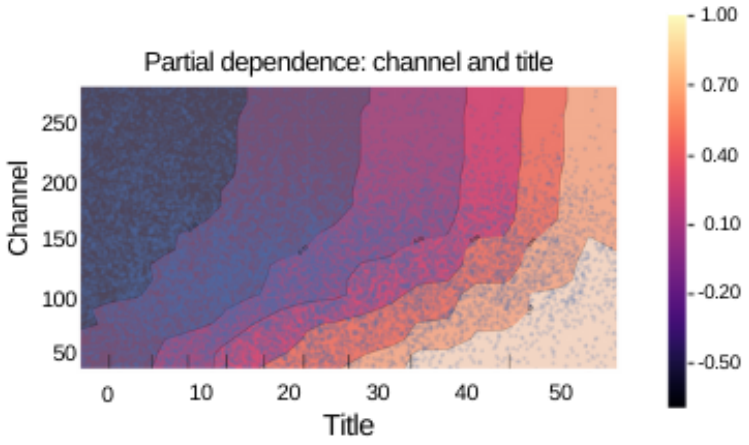


Figure 4.5: Two dimensional partial dependency map considering features channel and title. The blue dots shows the distribution.

One can observer a higher distribution in the upper left are of the image and how the distribution subsequently gets smaller as the blue dots becomes less and less dense as you go from the top left to the bottom right. There are almost no videos with a high title score and high channel score as can be seen in the upper right. The bottom right denotes a high title score and low channel score which yields the largest partial dependence value.

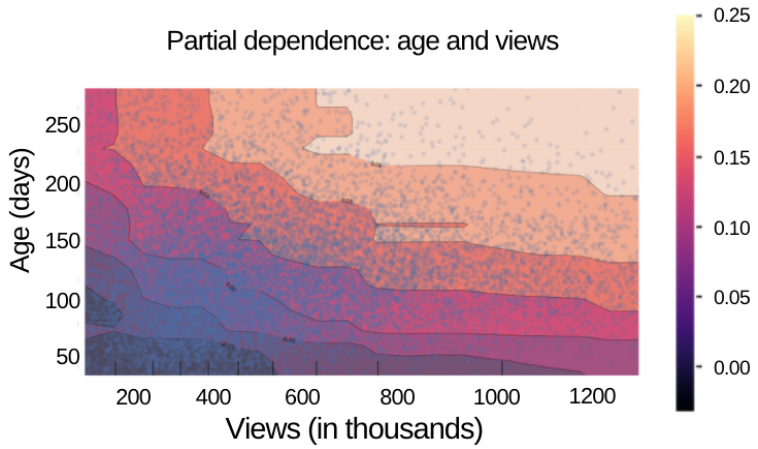


Figure 4.6: Two dimensional partial dependency map considering features age and views. The blue dots shows the distribution.

The map is very dense in the lower left area of figure 4.6 and becomes more and more spare as you go towards the upper right. While the upper right, high view and age, yields the highest partial dependence and has the most impact on video popularity, the amount of data with those values are very few. The overwhelming majority of the data has between 0-500 thousand views and are in the age range of 0-150 days.

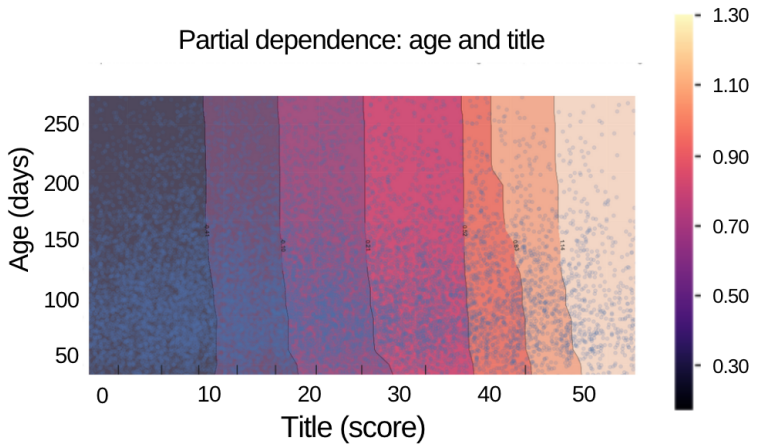


Figure 4.7: Two dimensional partial dependency map considering features age and title. The blue dots shows the distribution.

The distribution is even on the x-axis, but is a bit sparser in the higher age range. Age seems to have little impact video popularity when considering

the title. The value almost only varies depending on title score, changes in age does not affect it. A higher title score yields a higher partial dependence value.

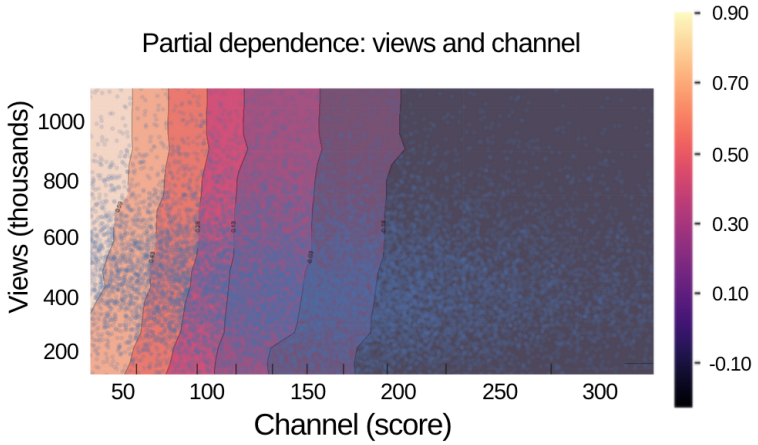


Figure 4.8: Two dimensional partial dependency map considering features views and channel. The blue dots shows the distribution.

The distribution is even in the center and smaller in the right area and the upper area where the views are higher. Similar to figure 4.7, The value almost only varies depending on channel score, changes in views barely affect the partial dependence. A lower channel score yields a higher partial dependence value and vice versa.

Chapter 5

Discussion

5.1 Discussion

There were three main objectives of the thesis which were stated in section 1.3. Those goals were to build a predictor model with two methods and evaluate and compare them, to extract feature importance to video popularity and lastly to measure how the subsequent features affected video popularity.

Knn had an average accuracy of 0.903 and ROC-AUC of 0.90 and logistic regression had an accuracy of 0.844 and ROC-AUC of 0.85. Knn performed slightly worse with higher k value. A higher k does not always imply better performance since a high k may result in under-fitting which explains table 4.1. The paper in section 2.5.3 also implemented a logistic regression methods for popularity prediction while adopting stepwise regression which yielded a superior result with an accuracy of 0.918%. Perhaps the accuracy can be further increased by utilizing other regression methods.

Figure 4.2 displays a diagram of feature importance. Features of most importance were channel, title, comment count, age and category in the mentioned order. Interestingly, the number of views had the second lowest importance. The paper in section 2.4.2 concluded that social features such as early views and comment count are the most important factors for video popularity. Perhaps early views are more integral to popular videos than total views. It is also feasibly that the low importance of the views feature is a result of YouTube's way of evaluating trending content, were they incorporate multiple data points to classify a video as trending or not. That would suggest that YouTube gives preference to user interaction data such as comments and dislikes rather than the sheer amount of views, which would explain why comment count and dislikes both had a higher importance than views.

Category was the most important feature together with description according to the paper in 2.4.1. While our results indicate that category is an important feature, there is a stark contrast on the importance of video description between our results and the paper. The authors implemented 4 different methods for predicting video popularity and choose extreme gradient boosting for their results which yielded the mentioned results. Perhaps the difference might lie in the different implementation and also feature selection since the selected features for their algorithms were title, time gap, category, tags, description, and duration while this thesis considered additional and different features.

Feature importance over different video categories was also considered. Of the 15 categories, 4 of them had a stark feature importance difference to the figure 4.3. The categories were music, pets & animals, sport and news & politics. All 4 categories put more emphasize on video title. Music, pets animals and news politics categories put more importance to the channel feature while sports category deemphasized the importance of the channel feature. Sports category put more concern on video description and music put less concern on the same feature. More research has to be conducted in this area to further rationalize feature importance for particular categories.

Figures 4.3 and 4.4 displays partial dependence for a single feature while marginalizing all other features, considering single feature as independent of the other features. The impact of the views feature increased as the number of views increased. Interestingly this was also the case for the video age. The older the video was, the more impact it had on video popularity. The dataset only contained videos during the year of 2018, one can observe that the graph is tapering of at 300 days. A larger dataset with videos spanning over multiple years would have given more insight to this dynamic. Similarly, figure 4.4 a) shows the partial dependence growing as the title score grew, however while the growth of figures 4.3 a) and b) started to subside for larger views and ages respectively, the partial dependence grew linearly with the title score. This is in contrast to the partial dependence of the channel feature which had a linearly inverse relationship. Bear in mind that title and channel score are based on tf-idf, a term frequency based scoring system. A higher tf-idf score denotes higher importance. However channel feature impact on popularity decreased as channel score increased. Perhaps scoring based on term frequency is not the optimal text to numerical conversion method in this experiment. A logical improvement would be to utilize multiple text to numerical conversion methods and further evaluate channel and title partial dependence.

Figures 4.5 to 4.8 shows multiple two dimensional partial dependence plot.

Contrary to the figures above, these figures considers two features as oppose to one. Two dimensional plots allows one to investigate how combinations of features affect the model output. Figure 4.5 displays the map with channel and title features. This plot concurs with plots 4.4 a) and b), where in both cases impact on popularity grew with the increase of title score and decrease of the channel score. However the distribution thins out with larger title. Figure 4.6 displays the 2D map with features age and views. The dynamic here in very much congruent to figure 4.3. The combination of higher views and more aged videos drives video popularity the most. Intuitively a more aged video allows for more views to be accumulated which would explain the explicit correlation, hence why there are few young videos (> 150 days) with higher views (< 700.000) and vice versa.

The subsequent two figures displays the map for features age, title and views, channel respectively. In both scenarios the impact for video popularity is relatively exclusive to the title and the channel. In fact, the dynamic is reminiscent of figure 4.4 where title and channel were considered independent. Perhaps this is a consequence of title and channel being the two most important features for video popularity, significantly more important than age and views.

5.2 Limitations

The thesis was limited in multiple ways. Only two models, knn and logistic regressions were implemented. Part of the discussion was to rationalize the results which was made more difficult since YouTube's method of evaluating trending content are not reported or publicly available. the general notion is that it is a combination of views, likes, comments, age where the video is coming from etc, but there could be additional components they consider.

Another limiting factor was the range of data. The data ranged from January 1st to December 31 2018. For instance, as previously mentioned, an increase in video age denoted increasing impact popularity until the increase rate slowed down after 300 days. Further observation could not be made since the data was maximum 365 days old. Observing partial dependence over a wider range of ages and other features could give more understanding to their impact on video popularity.

5.3 Future work

There are multiple ways to extend this work. Some logical improvements were mentioned in the discussion such as considering additional text to numerical conversion methods and working with datasets with larger range. Those suggestions are immediate development on this thesis. However, this work only considered popularity on YouTube and utilized data from one platform, without consideration of other popularity factors. There are multiple exterior factors for popular content such as real world events, and trending topics on the web which increases the difficulty of popularity content prediction and feature importance. Perhaps a good extension of this thesis would be a more holistic approach and consider the impact of additional factors outside of YouTube on YouTube video popularity. This could be further extended if one were to also consider real world events. Researching, purposing and improving methods for popularity prediction and feature extraction is also needed. There could also be more emphasis on the anatomy of a viral video. Predictive analysis is dependant on feature data however, what is it in the video that causes the popularity to spike? An approach dissecting the video structure and format would certainly aid in the analysis of what in fact constitutes a viral video.

Chapter 6

Conclusions

6.1 Conclusion

This thesis aimed to observe and further understand video popularity on YouTube. Two methods have been presented for predicting video popularity, logistic regression and k-nearest neighbour. These models have been used to classify popular videos and to extract features importance on video popularity while also observing relationship between the subsequent features and popularity impact. The knn method performed well, better than the logistic regression method.

The results suggest that channel, title were the most important features followed by comment count, age and category. The number of views was second to last in importance. Partial dependency figures were presented which shows the marginal impact of features on video popularity

This work can be extended in multiple ways and can be thought of as an experiment rather than a complete method and is not ready to be used in real world application. Many more exterior factors need to be considered for more accurate predictions and to more correctly analyze feature importance. There is as such, a need for further research in refining.

References

[1] W. Commons, "File:knnclassification.svg — wikimedia commons, the free media repository," 2021, [Online; accessed 28-May-2021]. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:KnnClassification.svg&oldid=547099866

[2] A. Internet, "The top 500 sites on the web," 2018. [Online]. Available: https://www.alexa.com/topsites

[3] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. Hauptmann, "Viral video style: A closer look at viral videos on youtube," 04 2014. doi: 10.1145/2578726.2578754

[4] K. Kumari and S. Yadav, "Linear regression analysis study," *Journal of the Practice of Cardiovascular Sciences*, vol. 4, p. 33, 01 2018. doi: 10.4103/jpcs.jpcs$_{81}$8

[5] A. Schneider, G. Hommel, and M. Blettner, "Linear regression analysis: part 14 of a series on evaluation of scientific publications," *Dtsch Arztebl Int*, vol. 107, no. 44, pp. 776–782, Nov 2010.

[6] T. G. Nick and K. M. Campbell, "Logistic regression," *Methods Mol Biol*, vol. 404, pp. 273–301, 2007.

[7] S. Sperandei, "Understanding logistic regression analysis," *Biochem Med (Zagreb)*, vol. 24, no. 1, pp. 12–18, 2014.

[8] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Ann Transl Med*, vol. 4, no. 11, p. 218, Jun 2016.

[9] P. Cunningham and S. Delany, "k-nearest neighbour classifiers," *Mult Classif Syst*, 04 2007.

[10] P. Nadkarni, "Chapter 10 - core technologies: Data mining and "big data"," in *Clinical Research Computing*, P. Nadkarni, Ed. Academic Press, 2016, pp. 187–204. ISBN 978-0-12-803130-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128031308000105

[11] H. A. Abu Alfeilat, A. B. A. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. Eyal Salman, and V. B. S. Prasath, "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review," *Big Data*, vol. 7, no. 4, pp. 221–248, 12 2019.

[12] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?" *Emerg Med J*, vol. 34, no. 6, pp. 357–359, Jun 2017.

[13] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," 2019.

[14] Y. Li, K. Eng, and L. Zhang, "Youtube videos prediction: Will this video be popular?" 2019.

[15] T. Trzcinski and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Transactions on Multimedia*, vol. PP, 10 2015. doi: 10.1109/TMM.2017.2695439

[16] S. Mekouar, N. Zrira, and E. Bouyakhf, "Popularity prediction of videos in youtube as case study: A regression analysis study," in *BDCA'17*, 2017.

[17] G. Szabó and B. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, 12 2008. doi: 10.2139/ssrn.1295610

TRITA -EECS-EX-2021:554