ST4248
Statistical Learning II

Term Paper

A0190036M

Summary

This project attempts to predict the Reception (like-dislike ratio) to a trending YouTube video based on its other features. The power of such a prediction lies in allowing content creators to assess the public reaction to their videos and making use of this information to correct or supplement future videos. The regression problem will seek to minimise the RMSE in tuning and deciding between a justified selection of models: Linear Regression and Random Forest Regression. The results will be analysed, and contextual links will be made to justify the observed patterns in the results presented in the report.

Introduction

In the internet age, sites like YouTube offer unparalleled power and reach in selling a product or promoting a cause to the masses. Equally important to having a wide support of subscribers is the quality of videos produced, with the reception to a video having the power to make or break a newly launched product. A natural indicator of this is the ratio of likes to dislikes, termed "reception" henceforth. This project aims to predict a trending YouTube video's reception using other information relating to the video and channel it is posted by. The applicability of this lies in the fact that the ability to predict a video's reception provides a basis for measuring whether a video is underperforming or overperforming its predicted like ratio, allowing content creators to take corrective measures if needed.

Dataset

The dataset used used consists of information on the 6254 trending videos on the main Youtube site (US region) from November 2017 to June 2018, sourced from Kaggle[1]. Restricting the analysis to trending videos not only provides a natural limit on the number of videos in the analysis, but also filters out low-interaction videos that form the majority of videos in Youtube and provide little to no discriminatory information (eg. due to having few likes and views) that would only realistically function as noise. Though, the possibility of expanding the analysis to other Youtube region and outside of trending videos will be discussed at the end of the report. The original dataset consists of the variables: Title, Publish date, Views, Likes, Comments and Category.

Feature engineering

The Likes and Dislikes variables are morphed into the response **Reception** variable by dividing the first by the second. The Category variable, originally in numbers (16 integers from 1-43) is coded into the category names (eg. Auto & Vehicles). The publish date of a video might not provide much information on its own due to a strong relation with the natural increase in views, likes, dislikes and comments for a video over time, and it is hence reengineered into a categorical variable indicating the day of the week the video is published. In addition, since the number of comments and number of views are generally expected to have a trivial positive relationship, a more meaningful interaction variable called **Engagement** is created that measures the number of comments per view, by dividing the number of comments by the number of views.

Since the Title variable is in the form of a text corpus, it is transformed into a numerical form that machine learning algorithms can easily extract information out from, and represent information in. Through research, it is found that Vader[2] and Textblob[3] are the two most comprehensive and well-maintained models for analysing the sentiment of a text. This project will institute the sentiment analysis Vader engine as it outputs values of greater precision than Textblob, allowing for more precise discrimination between the video titles, which is especially important considering **their short lengths compared to conventional text corpuses**. The extracted sentiment scores, which have an output range of -1 to 1, will provide a measure of discrimination between the expressions conveyed by the video titles. Table 1 provides the final variables used in the project.

[1] https://www.kaggle.com/datasnaek/youtube-new
[2] https://www.nltk.org/_modules/nltk/sentiment/vader.html
[3] https://textblob.readthedocs.io/en/dev/

Numerical	Categorical
• **Reception (Likes per Dislike)** • Comment Count • Views • Title sentiment • Engagement (Comments per view)	• Category • Weekday published

Table 1: Summary table of feature variables (response variable Reception is bolded)

Preliminary exploration

Some preliminary investigation is carried out to anticipate multicollinearity, with the results indicating that there are no alarmingly high VIF correlation scores for the features. Additionally, a pairwise scatterplot (Fig. 1) indicates no obvious visual relationships between the numerical features and the response.

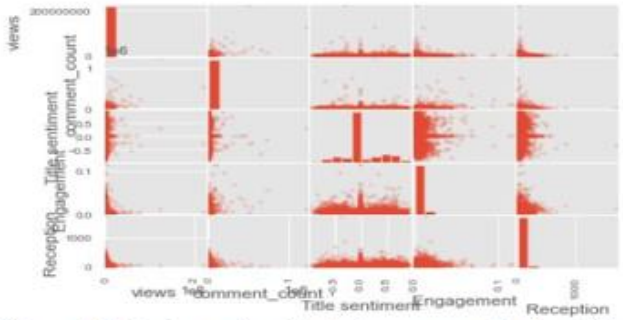

Figure 1: Pairwise scatterplots between numerical variables

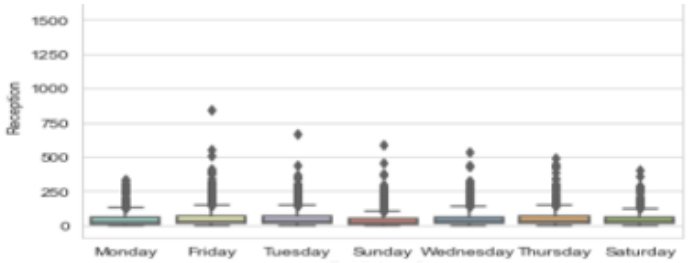

Figure 2: Boxplot for Weekday Published

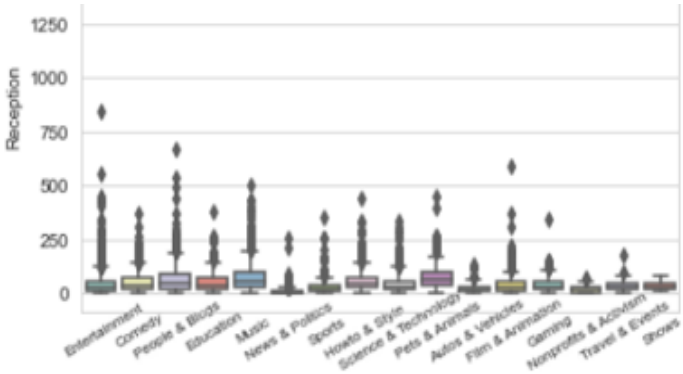

Figure 3: Boxplot for Category

Similarly, boxplots (Fig. 2) for the distribution of Reception for the different Weekday Published values indicates no significant difference between the weekdays, except for some outlier instances. However, a similar boxplot (Fig. 3) for Category shows differences in the distribution of Reception for different video categories, indicating that the Category might have a noticeable effect on a video's Reception. For instance, Music videos tend to have greater Reception than News & Politics videos, a phenomenon that could be contextually attributable to the nature of the videos in News & Politics having greater tendency to generate polarising effects especially between those with different political views, and hence attracting more dislikes leading to lower Reception.

Data splitting and scaling

For the analysis, the data is split into 75%-25% to be used for training and testing respectively. Additionally, standard-scaling was applied **separately** to the numerical independent variables of the training set, and then the test set. All training and cross-validation procedures for model fitting were performed using only the training set, before being evaluated on the training set to present the scores in this report.

Choice of assessment metrics

For this **regression** problem, the Root Mean Square Error (RMSE) will be used instead of the MAE (Mean Average Error) to weight larger errors more and tolerate some small errors. This metric will be used to tune the models with cross validation and present the impact of the tuned models on the training set. As such, the models will be optimised for best RMSE.

Models in consideration and explanation of model choices

A **linear regression** model will first be attempted with the associated tuning, regularisation, and feature selection procedures. After that, a **Random Forest regression** model will be tried to explore the possibility of relaxing linearity in the solution, and to provide a different approach to the regression problem. For instance, a tree-based mechanism like Random Forest is non-parametric and is also able to handle categorical data like that in the Weekday Published and Category features, something that linear regression might be disadvantaged with due to the need to dummy encode these variables which could inflate dimensionality too much, leading to overfitting. However, it is noted that this dataset consists of mostly numerical features, so it might not be unexpected to see this difference less pronounced, or even linear regression performing better in RMSE, an entirely plausible scenario owing to these two models having such different characteristics and mechanisms after all.

Model fitting, optimisation and results – Linear Regression

For the linear regression model, the categorical variables underwent n-1 one hot encoding into dummy variables first. An exhaustive 5-fold cross validation procedure is utilised over the entire train dataset to tune for a reasonable preliminary set of elastic net regression (to allow both L1 and L2 regularisation) hyperparameters for the combination with the best RMSE. The purpose of this is to find a reasonable preliminary set of hyperparameters that we can conduct the next step, feature selection, with. Using this set of preliminary parameters, a 5-fold cross validation recursive feature elimination (RFE) procedure is utilised to find the subset of features that gives the best RMSE score. This subset of features is then used to fit an elastic net regression model that uses hyperparameter tuning again to arrive at a final combination of hyperparameters with the best RMSE. The selection of hyperparameters tuned is given is Table 2, with the final hyperparameters in bold.

Hyperparameter	Values tested
Fit Intercept	**True**, False
Normalise features	True, **False**
Alpha (regularisation strength)	**0.00005**, 0.0005, 0.001, 0.01, 0.05, 1, 2, 5, 20, 50, 100
Ratio of L1 to L2 regularisation	0.1, 0.2, 0.3, 0.4, **0.5**, 0.6, 0.7, 0.8, 1

Table 2: Tuned hyperparameters for Linear Regression (the final optimal values found are underlined and in bold)

Take note that the low final regularisation could mean that the feature selection has already done a decent job at mitigating overfitting. The final linear regression model has the following features and coefficients:

Feature	views	Engagement	category_Comedy	category_Education	category_Entertainment	category_Film & Animation	category_Gaming	category_Howto & Style
Coefficient	-2.558473	8.768864	17.399	20.52331	12.55971	9.127424	10.67147	21.86415

category_Music	category_News & Politics	category_Nonprofits & Activism	category_People & Blogs	category_Pets & Animals	category_Science & Technology	Published On_Monday	Published On_Saturday
45.92842	-25.11584	7.348451	28.18402	51.94679	16.58964	-6.350555	-5.014682

This final model and parameter set is then fit on the entire train set, and then evaluated on the test set, giving a RMSE of 55.963.

Model fitting, optimisation and results – Bootstrap Random Forest

Fit only on the training data, an exhaustive 5-fold cross validation grid search was conducted over a few important parameters:

Hyperparameter	Values tested
Number of estimators	10, **50**, 100, 500, 1000
Max. number of features to sample from at each split	n, **log(n)**, sqrt(n) *where n = total no. of features*
Bootstrap	**True**, False
Max. depth	None, 1, 2, 3, 4, 5, **10**, 25
Min. number of samples for split	2, **3**, 5, 10

Table 3: Tuned hyperparameters for Random Forest (the optimal values found are underlined and in bold)

This final model and parameter set is then fit on the entire train set, and then evaluated on the test set, giving a RMSE of 57.048..

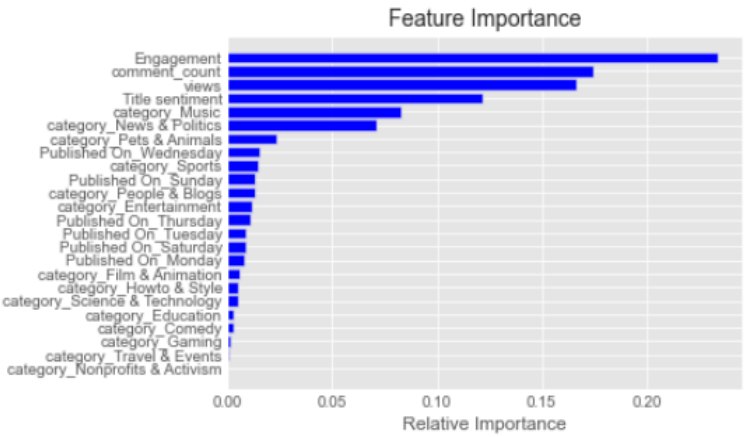

Figure 4: Feature importance for Random Forest model

The feature importance plot (Fig. 4) indicates that Engagement, comment count, number of views, Title sentiment, and being "Music" or "News and Politics" videos are the factors that affect a trending video's Reception the most, according to the tuned random forest model. On a contextual basis, these are quite possibly explainable. For instance, with regards to the importance of Engagement, how engaging or controversial a video is will easily affect its ratio of likes and dislikes. In addition, the count of views or comments is an indication of the spread or outreach of a video, and videos that have outreaches far out of the main/usual subscriber base of a channel might invite viewers or commenters less interested in the content sector/domain of the video or channel, who will have different Reception towards a video than the channel's usual viewer base. The relation of title sentiment to the Reception of a video is less straightforward, though a plausible explanation could be of the title setting the first impression of a video to viewers, who then judge the video based on whether it lived up to or matched their expectations created by the video title. Unsurprisingly, being a Music video or a News and Politics video has significant sway on the Reception to a video, as explored in the preliminary exploration segment of this report (Fig. 2).

Analysis and comparison of results

As seen from the summary table (Table 4), the Linear Regression model performed better than the Random Forest model in terms of RMSE. One reason for this could be that the numerical features are the key features predicting the Reception, a suggestion also supported by the fact that the

	Linear (Elastic Net) Regression	Random Forest
RMSE	**55.963**	57.048

Table 4: Summary Table

features Engagement, comment count, number of views, and title sentiment dominated the Random Forest feature importance chart discussed in Fig. 4, although title sentiment and comment count did not make the final cut based on the RMSE-minimising RFE procedure for Linear Regression.

However, the difference in performance is not extreme and perhaps if the sample size were larger, this could provide more support especially for the 16-class feature 'Category'. In theory, tree-based methods like Random Forest might, due to the ability of the splitting mechanism to directly deal with categorical variables, have a dimensionality advantage over linear regression if a larger proportion of the dataset is of such variables, which is not the case here. Although dummy encoding the Category feature increased dimensionality for the linear regression model, any potential overfitting could have been mitigated in part by the feature selection and regularisation procedures and hence this did not disadvantage the model (over Random Forest) as previously feared. As such, if more of the features were categorical, the results could very possibly have been the opposite.

All in all, evidence indicates that a linear boundary like the one in Linear Regression might be sufficient for a decent fit for Reception based on these features, without needing to introduce nonlinearity, say, through a Random Forest model.

Conclusion

In conclusion, this project explored regression using Linear Regression and Random Forest for predicting Reception to a trending YouTube video. It is conclusive that the Linear Regression model was the best performing model, based on both RMSE and MAE. Using this model, content creators can predict the expected Reception towards a video based on its other features, and compare this to the actual Reception to assess if a video is "overliked" or "underliked", and hence repeat similar styles and themes for overperforming videos in future, or correct less popular characteristics or flavours in "underliked" videos.

Although the project utilised Vader for sentiment analysis, an approach like the one used here may hold even more promise in future with the development of even more precise sentiment analysis methods, particularly because of the relatively short length of YouTube video titles as compared to most conventional corpuses, thus requiring information be used efficiently down to the letter. This would of course be contingent on having a sufficient sample size to support the increased precision, which could be achieved by extending the geographical and temporal scope of the YouTube data used.

On this note, future extensions of this project can also explore different geographical regions other than YouTube US, or possibly investigate the differences in the relationship strength between Reception and the features for different regions, or even extending the scope to non-trending videos. These could possibly bring different insights due to the vastly different characteristics (eg. Engagement) between trending and non-trending videos, and between trending videos in different countries (eg. 'Music' category videos might dominate the Trending list in one country, and 'People & Blogs' for another country).