DS 310 In-Class Project 1 Report

Kaggle Team name : jecseB

Members : Ben Giles and Jesse Bao

EDA/Data pre-processing

Imported libraries: We chose to import the sklearn, scipy, numpy, and other libraries because we felt like they would have useful components for our project.

We explored the dataset a bit to understand the features we were given to predict clickbait videos. We took a look at the most popular words(strings) found in clickbait and non-clickbait headlines. We checked the dataset to see if there were any NA values that we had to deal with and changed the heading of the clickbait/non-clickbait column as well as the type to int. We also checked the correlations between the clickbait/non-clickbait column with the other numerical columns.(i.e. likeCount, dislikeCount, etc.) We created a new dataframe containing the title of the video and the column indicating clickbait as a separate data frame we could use to check our predictions later.

Feature Engineering

We thought that terms like exclamation marks, question marks, words that signified questions(who, what, when, where, why, etc.), http signifying links, and numbers in the video title and descriptions might be good indicators of clickbait so we wrote functions to create columns determining whether video entries had these in their data. Additionally, we thought comment and description sentiment analysis would be helpful in predicting the data so we tried using some basic text analysis tools and it improved our model.

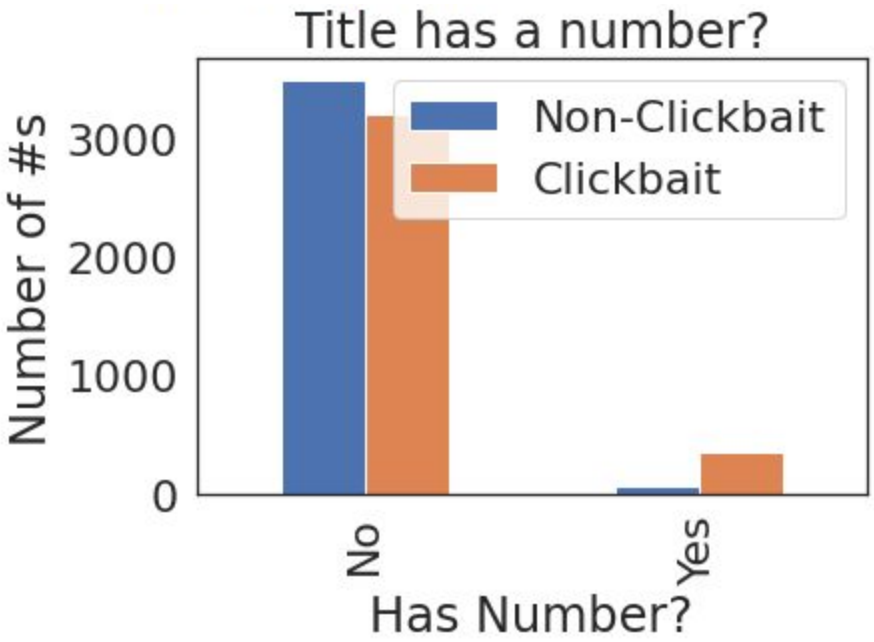

Chart showing if the title has a number for clickbait and non-clickbait videos

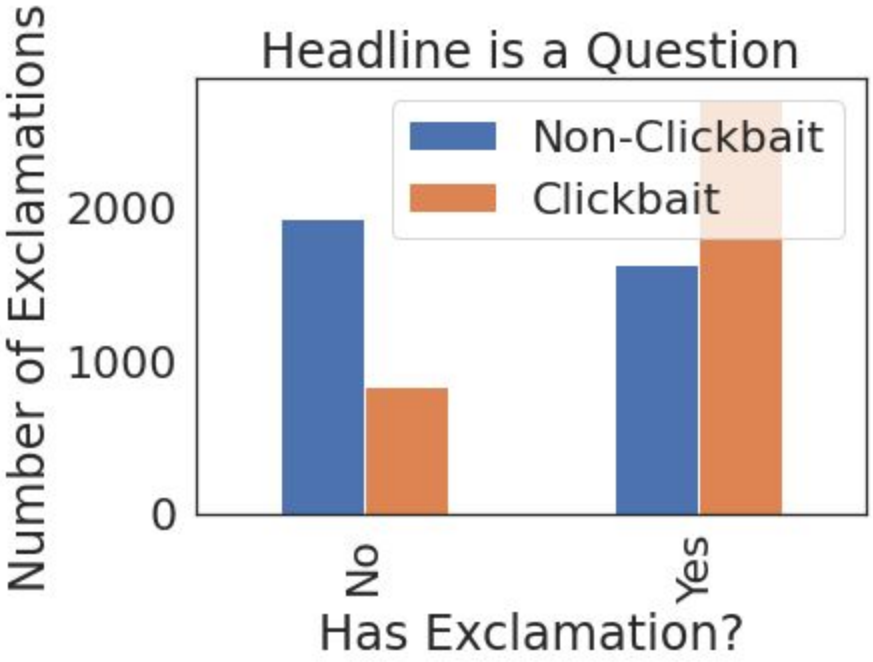

Chart showing if the title has an exclamation point for clickbait and non-clickbait videos

Explanation of the chosen ML model with rationale

Logistic Regression and Naïve Bayes were the first two types of classifiers that came to mind when it came to picking a model to use for this project. We related finding clickbait to finding spam in Naïve Bayes especially if there turned out to be common words in clickbait videos. We also thought that the problem was similar to the logistic regression problem we did in a previous homework assignment so we decided to build models for both types of classifier to see which one was better specifically for our situation. Our logistic regression model performed far better than our Naïve Bayes in terms of accuracy which achieved us an initial submission value of 0.70 F1 score. We ended up trying out a RandomForest model on the data afterwards to see how it would perform compared to our logistic regression and Naïve Bayes model and this provided us an even better result than logistic regression so we ended up settling on the Random Forest model.

GaussianNB classifier F1 score

```
[ ]  gnb = GaussianNB()
     y_pred2 = gnb.fit(X_train, y_train).predict(X_test)
```

```
[ ]  print('Accuracy of RF Classifier:', model2.score(X_test, y_test))
     print("F1 score:", f1_score(y_test, y_pred2, average='binary'))

     Accuracy of RF Classifier: 0.9479239971850809
     F1 score: 0.3672131147540984
```

Any hyperparameter setting (if any):

We did not tune any hyperparameters.

Performance evaluation in training

Our initial goal was to hit the baseline score, because we could always go back and improve our model afterwards so we started with very little engineered features and went

back searching for more possible correlations after viewing the results of the model with the features at that time.

Best testing F1 score in training

```
print('Accuracy of RF Classifier:', model2.score(X_test, y_test))
print("F1 score:", f1_score(y_test, y_pred, average='binary'))
```

```
Accuracy of RF Classifier: 0.9479239971850809
F1 score: 0.9491758241758242
```

How this can be used practically:

This model could easily have a use in sorting youtube videos that popup in your search recommendations. By filtering out clickbait videos you would get more quality videos in the topic you searched for. Our model is currently around 91% accurate as is and that might be an acceptable threshold for some who are willing to lose a couple good videos to completely get rid of clickbait. However if you really do not want to lose any non click bait videos this accuracy might be a bit low.

Screenshot of leaderboard