Final Report:
Mashable's Shares Analysis

Problem Statement

The internet has become a huge source where the majority of people can go for their news and entertainment. There are a lot of other sites where people can go, which creates a lot of competition. Mashable is one of the millions of websites where people can get all their news and entertainment from. Mashable is an international entertainment, culture, tech, science and social good digital media platform, news website and multi-platform media and entertainment company.

Here we are trying to attract more clicks/viewers to Mashables by analyzing which category of articles is popular. With that information, are we able to use this to focus and display this popular category to attract more clicks and attention?

We will use machine learning techniques in order to figure this out and predict whether or not the articles will be popular before being published. The number of shares will represent the popularity of the article.

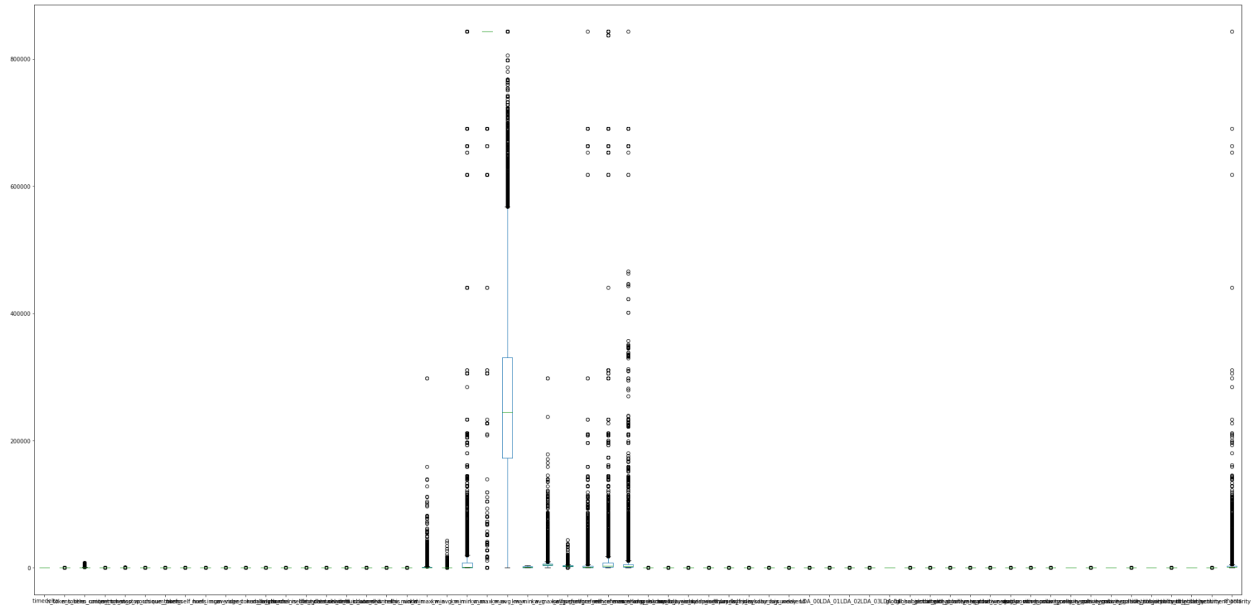
Data Wrangling

The raw data was obtained from the OnlineNewsPopularity.csv which can be downloaded from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#). The data contained 39644 rows and 61 columns, the 61 columns also known as the variables can be seen through the UCI machine learning repository portal. We looked into the data and saw if there are any missing values. In this case we were lucky to see that there were no missing values in the data.

Exploratory Data Analysis

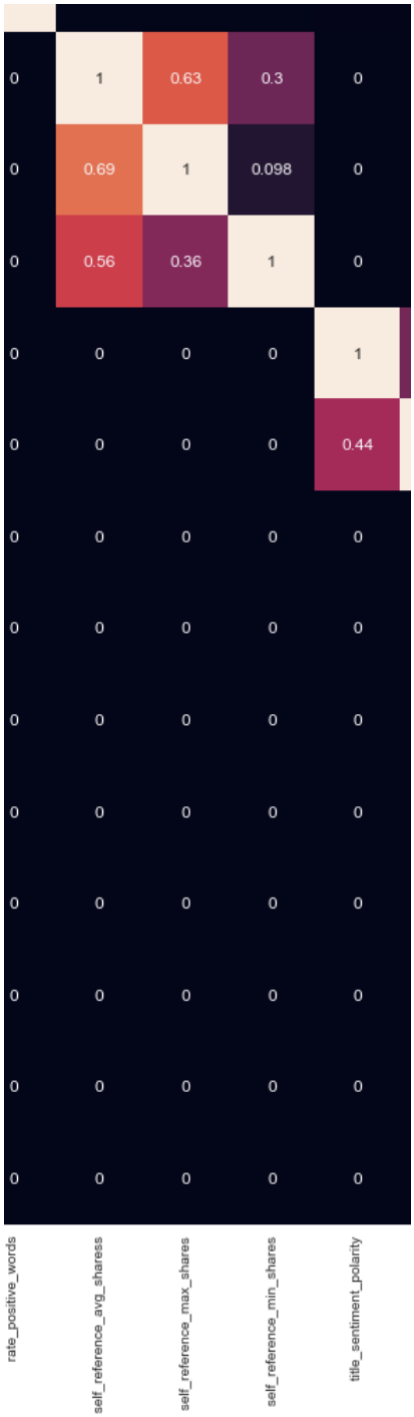
In order to create a good and an accurate model, we first take a look and see if there are any outliers in the data. A box plot was created to visualized any outliers in this data as we can

see in this figure:



Looking at this figure we can see that there are outliers from the data. We then look into it clearly by plotting each individual columns' boxplots. There were many outliers as we can see through our analysis and we wanted to remove the outliers since they can create a problem with our model. We eliminated as many outliers as possible but not all of them using the IQR (interquartile range) method.

Heatmap was created using seaborn to show any correlations between the variables. There are several variables that have high correlations for example as we can see from the figure below:



:

This meant that they provided similar information that could be repetitive during the prediction of a target variable. Here we can see examples of self_reference_avg_shares, self_reference_max_shares and self_reference_min_shares whose correlations are close to 1.

Preprocessing, Training and Modeling

To begin creating our model we opted to use the classification module instead of the regression module for an accurate model. First we went ahead and preprocessed the data as it was taking too long for it to be processed. To do this we processed any popular shares that are equal to and greater than 1400 and unpopular shreas less than 1400. We set and renamed the popular shares as 1 and the unpopular shares as 0. We filtered out the popular shares to be processed.

In order to create a model we had to set up the dataset. We used the default test split of 70/30 in this experiment.

4	Original Data	(35103, 58)
5	Missing Values	True
6	Numeric Features	36
7	Categorical Features	21
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None
11	Transformed Train Set	(24572, 90)
12	Transformed Test Set	(10531, 90)

This figure here shows the original shape of the dataset is (35103, 58), which meant that there were 35103 samples and 58 features including our target column (shares). The transformed train set was (24572, 90), the features increased due to categorical encoding. While our transformed test set from the 70/30 split came out to be (10531, 90).

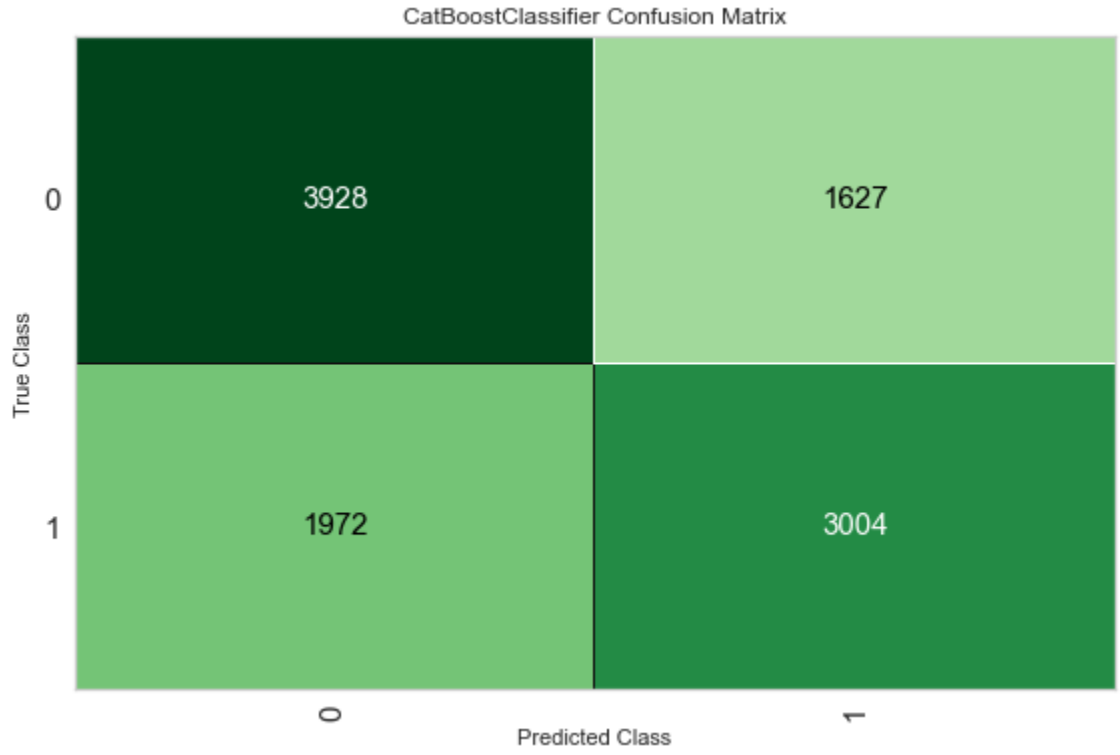
We compared the model to select the best one to use in this.

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6591	0.7184	0.6151	0.6474	0.6308	0.3145	0.3149
1	0.6615	0.7198	0.5902	0.6593	0.6228	0.3175	0.3191
2	0.6659	0.7258	0.6053	0.6604	0.6317	0.3270	0.3280
3	0.6695	0.7365	0.6242	0.6594	0.6413	0.3354	0.3358
4	0.6671	0.7344	0.6053	0.6623	0.6325	0.3293	0.3304
5	0.6459	0.7093	0.6040	0.6322	0.6178	0.2883	0.2885
6	0.6577	0.7173	0.5997	0.6505	0.6241	0.3108	0.3117
7	0.6748	0.7372	0.6289	0.6661	0.6469	0.3460	0.3465
8	0.6654	0.7265	0.6237	0.6541	0.6385	0.3275	0.3278
9	0.6768	0.7307	0.6418	0.6646	0.6530	0.3508	0.3510
Mean	0.6644	0.7256	0.6138	0.6556	0.6339	0.3247	0.3254

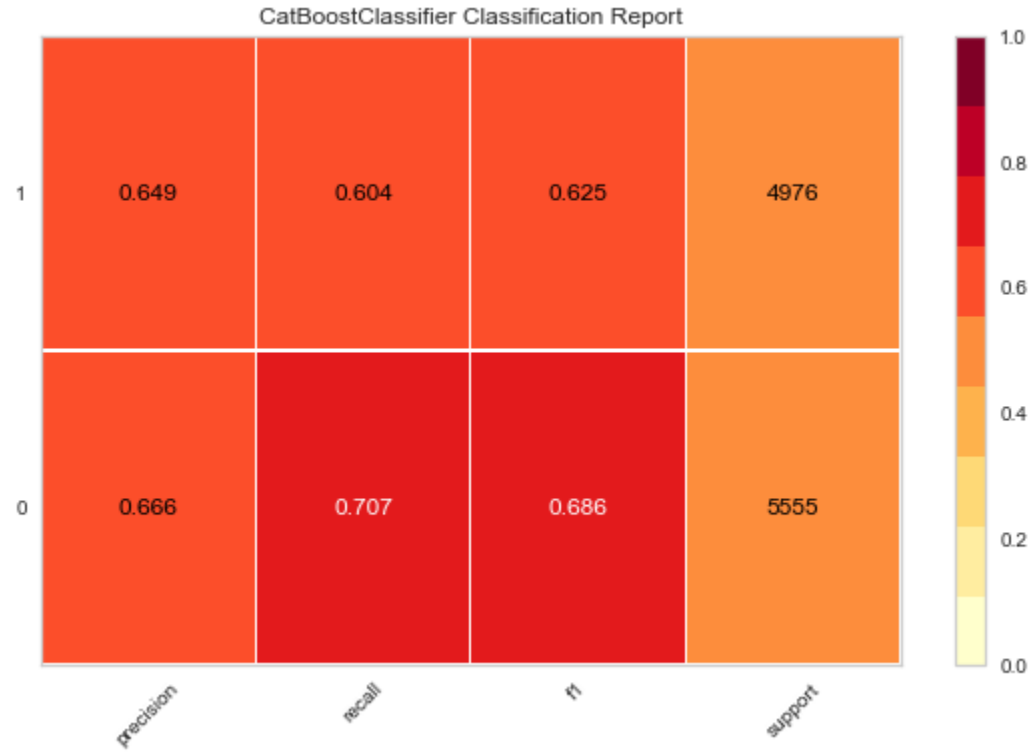
We decided to use the CatBoost Classifier model as it shows the highest accuracy of 0.6644 compared to the other models. After tuning the hyperparameters of the model, the final results shows:

Mean	0.6647	0.7244	0.6064	0.6587	0.6313	0.3248	0.3258

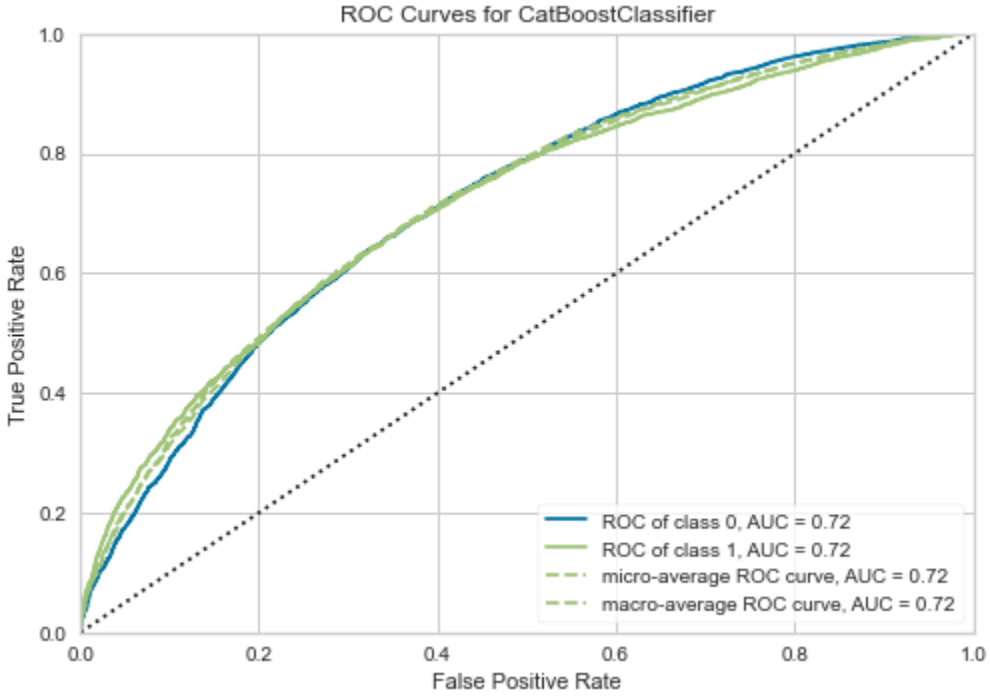
Models

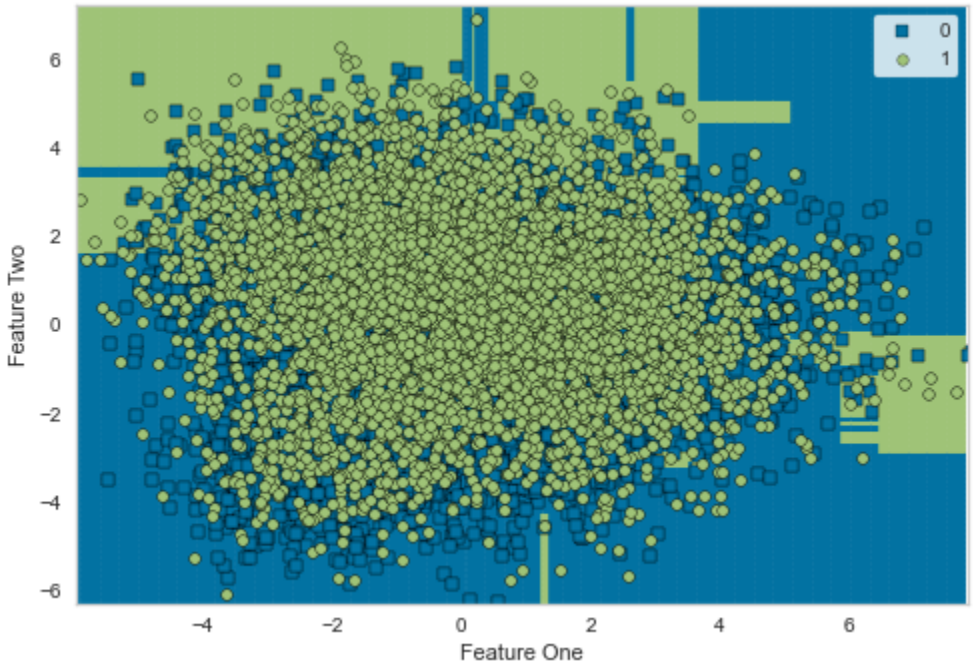


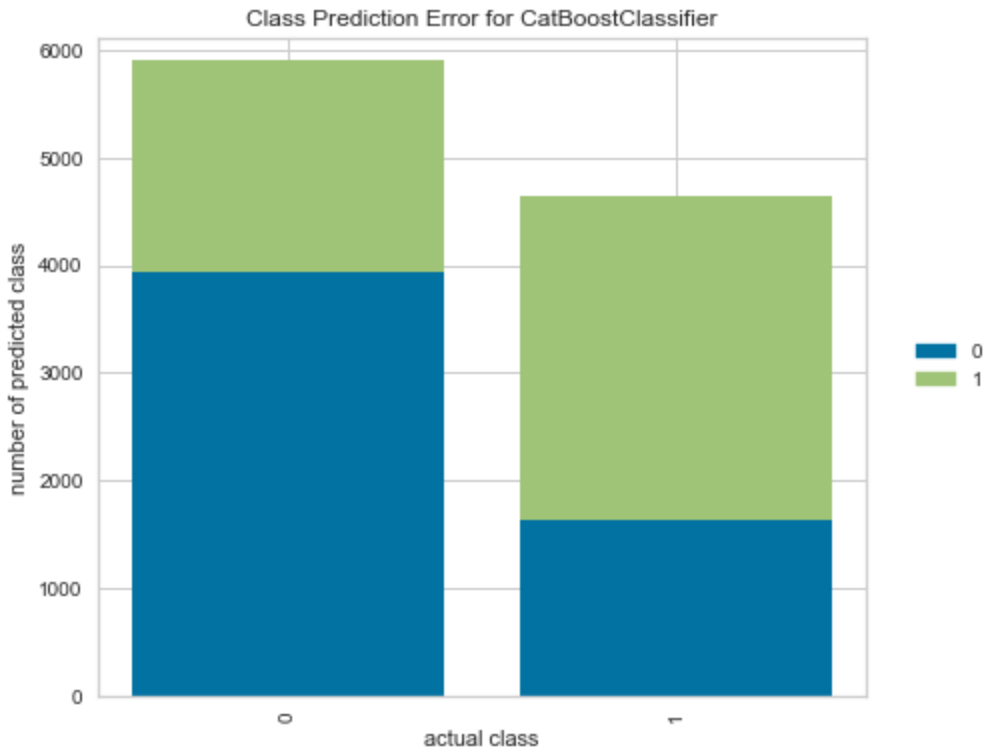
In this confusion matrix the algorithm accurately predicted 3928 unpopular shares and 3004 popular shares.



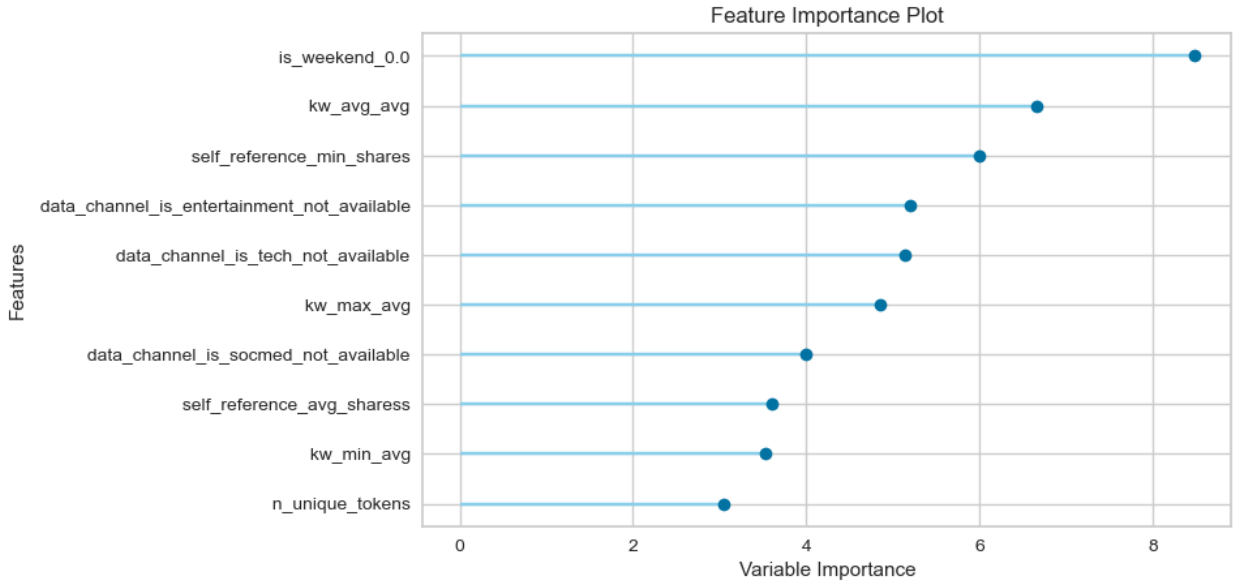
To further more to see how accurately our model predicted which we can see here from the classification report. The percentage of both the popular and unpopular shares shown above 50 percent, around the 60 percent.







These three figures are more visualization we can use to identify the accuracy of our model.



This figure is an important plot for this experiment as it explores the importance of our variables. This is showing that the popular number of shared articles come from the weekend. It also shows us that the categories of articles in entertainment, tech and social media have the greatest number of shares compared to the others. Lastly we can see that unique tokens also known as popular keywords are relevant in contributing the number of shares.

Take Away

Based on the models we created and the accuracy of about 66 percent. We can see from the feature importance plot that to attract more viewers we can increase the articles in the tech, entertainment and social media subjects. Doing so will help us increase the numbers of shares. We can also see that having unique keywords we can call popular keywords are crucial to having popular articles. Also, the amount of popularly shared articles tend to be shared on the weekend.

Our recommendation is to increase the amount of articles to be published in the categories of tech, entertainment, and social media. We also can recommend having to improve the keyword of the articles to attract more viewers. We can have a team that can concentrate and dedicate to that task. For all other articles we can publish it during the weekend to reach a larger audience, since weekends seem to have the most popular shares.

Future Improvements

In the future we can try a more advanced cross validation even though this may take a longer training time. Also, instead of preprocessing the model we can process the entire dataset from the categorical module. We can also focus more on the articles themselves and create a model on the categories of the articles with the popular keywords. If we have more data on the popular keyword we can create a model and see which words are popular. In this case we can have the writers use those keywords to see if we attract more views in their articles.