Online News Predictor

The Data

➔ The dataset summarizes a set of features about articles published by Mashable , a news website over a period of two years.

➔ Data Source : UCI ML Repository (https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity)

➔ Number of attributes : 61

➔ Number of records : 39645

➔ Target variable : Number of shares

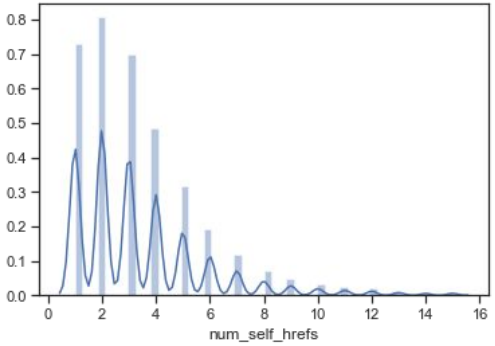
➔ No missing values

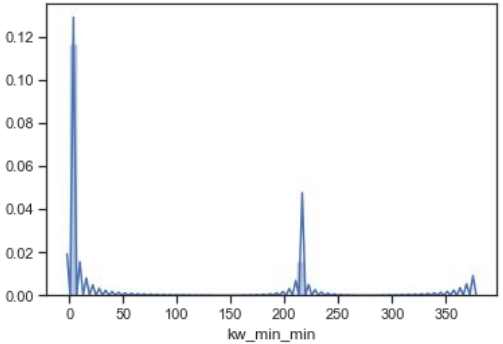
➔ The objective is to predict if an article will be popular or not

EDA

The maximum number of shares articles got were in the range 1000 to 3000.

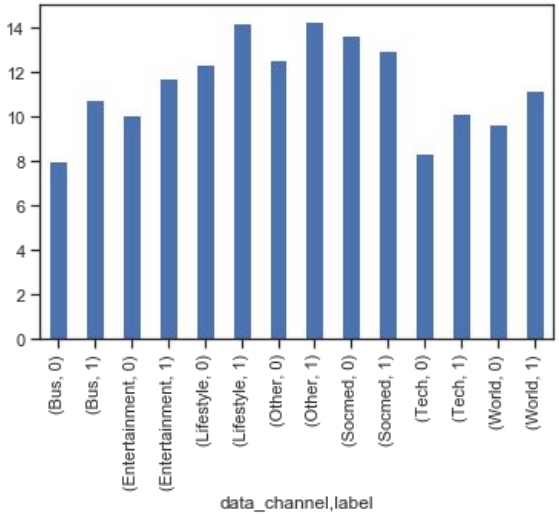
IQR and Z-Score method to remove outliers.





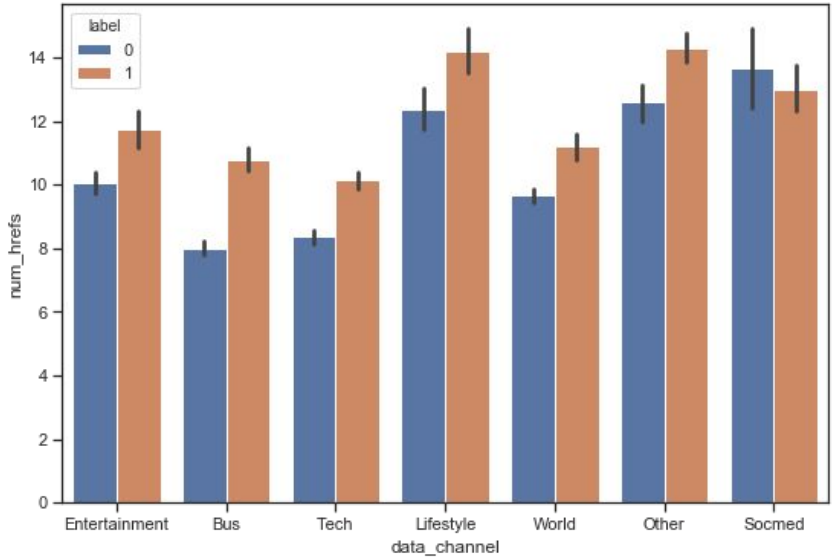
Outcomes

Lifestyle data channel got the maximum number of shares. While social media got the least number of shares.



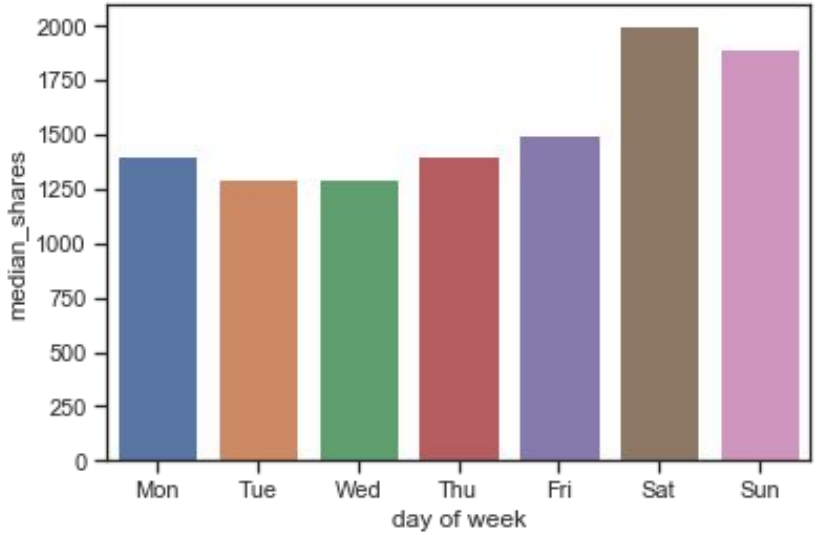
Popular Topic

Most popular data channel was Lifestyle followed by social media.



Popular Day

So the best day to publish an article can be Saturday.

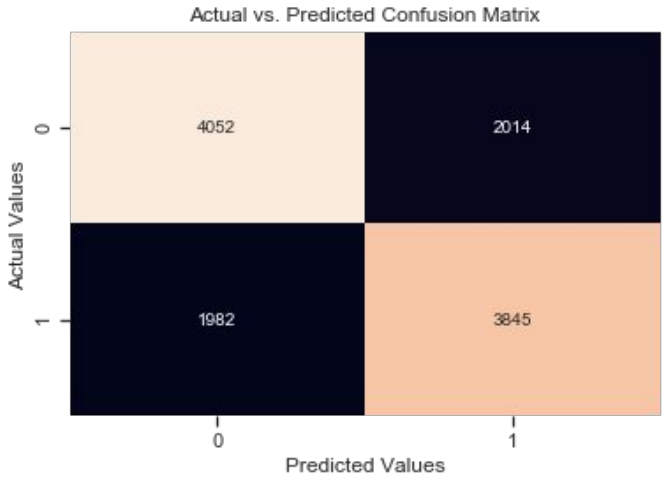


CLASSIFICATION RESULT:

Model	Accuracy	Precison	Recall	F1_score
Logistic Regression	0.639887	0.6017	0.600668	0.6002
Logistic Regression with vif	0.680528	0.6407	0.632413	0.6320
Decision Tree with gini criterion	0.653821	0.6041	0.620860	0.6201
Random Forest	0.730982	0.6566	0.665626	0.6656
xgboost	0.730821	0.6537	0.663213	0.6631

- The highest accuracy achieved was from random forest
- XGBoost gave an accuracy near to random forest but the precision and recall scores are more for RF.

Random Forest: Confusion Matrix



- True positives: 3877
- True Negative: 4040
- False Positive: 2028
- False Negative: 1949
- Precision=TP/ (TP+FP) =3877/ (3877+2028) =0.6566
- Recall=TP/ (TP+FN) =3877/ (3877+1949) =0.6654

Recommendations

- Publish new articles during the weekend so that they become more popular
- Articles are getting more number of shares when the data channel is either lifestyle or social media
- Short and to the point: the articles with the most number of shares should have titles that are around 6 to 14 words.

Future Work

- Delve more into subtopics
  - For example: different topics under “lifestyle”
- Reasons why some topics are not as popular
  - This would be fun to partner with a psychology expert

Thank You