SmallSEOTools

PLAGIARISM SCAN REPORT

Words	1000	Date	April 05,2021
Characters	6300	Excluded URL	

0%	100%	0	49
Plagiarism	Unique	Plagiarized Sentences	Unique Sentences

Content Checked For Plagiarism

Firstly, we run RFECV with a logistic regression estimator. The cross validation score versus the number of feature selected is shown in Fig. 5. From the figure, we can find that there is a drop of score when number of features is 29. Thus RFECV algorithm select 29 most relevant features from 58 original features. The selected 29 features are listed in Fig. 6. Interestingly, the day of week and article category features are included in these 29 features.
Next, we run RFECV with a RF estimator. The cross validation score versus the number of feature selected is shown in Fig. 7. We find that RFECV selects 56 features for RF, which is almost the full features in dataset. The only two features RFECV excludes for RF are [' n non stop words', ' data channel is lifestyle'].
Lastly, we run RFECV with Adaboost estimator. The cross validation score versus the number of feature selected is shown in Fig. 7. From the figure, we can find there is a peak when number of features is 18. Thus RFECV algorithm select 18 most relevant features from 58 original features.Before algorithm implementation, for each algorithm, We also randomly split the dataset with its own selected features into training set (90%) and testing set (10%). The logistic regression, RF and Adaboost are implemented by the sklearn function, LogisticRegression(), RandomForestClassifier() and AdaBoostClassifier(), respectively. At this stage, We will first use the default setting for model hyperparameters. The default hyperparameters are: logistic regression - {"C": 1.0}; RF-{"n estimators": 10}; Adaboost-{"n estimators": 50, "learning rate": 1.0}. We try to refine the parameters in next part. With the selected feature for each algorithm, We run the three classification algorithms and their performance is shown in Fig.
The three metrics (accuracy, F1-score and AUC) are summarized in Fig
9. Under default parameter setting, Adaboost performs best in all three metrics, RF performs better than logistic regression in AUC while logistic regression performs better than RF in accuracy and F1-score. As for the training and testing speed, logistic regression is much faster than the other two, and RF runs faster than Adaboost.
After initial implementation and further refinement for the three classifiers, we find the best performance is obtained by the RF classifier with 500 trees in the forest. The best obtained metrics of RF has the accuracy of 0.6769, F1-score 0.7073 and AUC 0.6734. The final scores are not exceptional, which is sort of within the expectation, because the dataset is not linear separable as shown in the PCA in Section 2.1. But it still achieves a reasonable performance in news popularity prediction compared with a random guess. To test the robustness of the model, we change the split ratio of training/testing set from 0.1 to 0.15, and then run the three classifier with the same refined hyperparameters. Now the metrics are shown in Fig. 10. Compared with Fig. 9, the performance of the model is still similar and the best performance is still given by RF.We come to conclusion after comparing the results obtained from all the three classifiers used that Random forest algorithm proves to be the most accurate amongst all giving us an accuracy rate of 67%. The training and testing time of RF classifier is greatly increased since 500 trees are used in the forest, but it helps RF achieve the best performance in terms of accuracy, F1-score and AUC.As mentioned in Section 2.1, some data preprocessing works have been done by the data's donator. The categorical features like the published day of the week and article category have been transformed by one-hot encoding scheme, and the skewed feature like number of words in the article has been log-transformed. Based on this, We further preprocess the dataset by normalizing the numerical feature to the interval [0, 1] such that each feature is treated equally when applying supervised learning. We also select the median of target attribute as the threshold to convert the continuous target attribute to boolean label. Since there are 58 features in the dataset, it is reasonable to conduct a feature selection to reduce the data noise and increase the algorithm's running speed. One effective way is using recursive feature elimination with cross validation (RFECV) to automatically select the most significant features for certain classifier. Sklearn provides a function called

REFCV() that can help us.

Next, we do the principle component analysis (PCA) to visualize the data. As shown in Fig. 4, we project the data point onto first 2(4.a) and 3(4.b) principle components, respectively. It is clear that the dataset is not linearly separable in PCA space.

By observing various features, We think there are several relevant features like day of the week and article category. In Fig. 2, the count of popular/unpopular news over different days of the week is plotted. We can clearly find that the articles published over the weekends have larger potential to be popular. It makes sense because it is very likely that people will spend more time online browsing the news over the weekends. In Fig. 3, the count of popular/unpopular news over different article category is plotted. We can observe that in category of technology (”data channel is tech”) and social media (”data channel is socmed(social media”), the proportion of popular news is much larger the unpopular ones, and in category of world (”data channel is world”) and entertainment (”data channel is entertainment”), the proportion of unpopular news is larger than popular ones. This might reflect that the readers of Mashable prefer the channel of technology and social media much over the channel of world and entertainment.The dataset consists of 39,643 news articles from an online news website called Mashable collected over 2 years from the time period of Jan. 2013 to Jan. 2015. It is downloaded from UCI Machine Learning Repository as https://archive.ics.uci.edu/ml/datasets/Online+News+ Popularity# and this dataset is generously denoted by the author of [4]. For each instance of the dataset, it has 61 attributes which includes 1 target attribute (number of shares),

Sources	Similarity