Predicting Popularity of Online News Articles

Swapnil Deshpande

1.Domain Background

With the advent of the Internet news articles are shared multiple times within moments they published. Also, they often go beyond the published location and being shared worldwide. It is important for the company to see which articles are popular because that can influence their editorial decisions and shows what readers like to read from their publication. For ad supported news model, this provides insights into which articles will fetch more ad revenue.

Lot of methods are used for promoting articles including but not limited to sharing on social media. To make news popular, catchy headlines are often used in predatory way but, there is no reliable way to understand the future popularity of the article. With more focus on removing fake news or misleading article titles and increasing journalistic standards, identifying subject areas and articles which lead to popularity of news article and media house is financially important for publishers.

Academicians have done research on prediction of online news popularity. They have used multiple different criteria to decide if article is popular or not. Tatar Alexandru et al. [1] used comments on article as measure of popularity, (Hensinger, Elena, Ilias Flaounas, and Nello Cristianini) [2] defines popularity in terms of competition where popular articles were the most appealing on that day. Support Vector Machine (SVM) is used to classify most appealing / non-appealing articles of day. In this project we will use dataset of around 40,000 articles from *Mashable*. We will try to find best classification algorithm to predict if news article will become popular or not, before its publication.

2.Problem Statement

In this project, our aim is to use machine learning algorithms to predict article will be popular or not, prior to its publication. This is binary classification problem. The popularity is measured in

number of shares here, if number of shares are higher than threshold then article is popular otherwise not.

3.Dataset and Inputs

For this project we will dataset provided by *Mashable*, which includes 39797 articles published in a period of 2 years. This is publicly available dataset can be downloaded from https://archive.ics.uci.edu/ml/datasets/online+news+popularity

There are 61 variables available in the dataset with target variables is number of shares. We will convert target variable in the binary variables of popular or not popular using appropriate threshold on number of shares. Removing non predictable variables such as URL of the news article and number of the day between article publication and dataset acquisition, we are left with 58 variables.

4. Solution Statement

The solution approach to this problem is to apply classification algorithm to classify news articles into two categories namely popular and non-popular based on the number of shares. First, we will identify appropriate threshold on number of shares to decide if article is popular or not.

In next step we will split the data into test and train dataset and apply classification algorithm, Since, it is binary classification, we will use logistic regression and ensemble methods along with other classification algorithms.

The evaluation metrics proposed to use will be described below.

5. Benchmark Model

Fernandes, Vinagre and Cortez [3] donated this dataset and published their research. They used multiple models and received best results using Random Forest Classifier with accuracy of 0.67, F1 score 0.69 and Area Under Curve (AUC) 0.73. We will build the model which will either improve or closely match the performance of this model.

6. Evaluation Metrics

We will use Accuracy, F1 score and Area Under Curve to measure the performance of our model against the benchmark model.

a. Accuracy: Accuracy is the ratio of correct prediction to total predictions

$$Accuracy = \frac{true\ positives + true\ negatives}{total\ predictions}$$

b. F1 Score: F1 score is harmonic mean of precision and recall. We consider it robust measurement since it is not dependent on data distribution.

$$F1\ Score = \frac{2 * precision * recall}{precison + recall}$$

c. AUC is Area Under ROC Curve which is plot of true positive rate vs false positive rate.

7. Project Design

Data Preprocessing

1. Check for missing values, outliers in the data
2. Do the exploratory analysis to understand the data
3. Convert number of shares in binary variable
4. Convert the data types in the appropriate form for the classifier algorithm
5. Split the data into test and training dataset

Model Training

1. Starting with the simple model check various classifier algorithm such as Logistic Regression, Naïve Bayes, Decision Trees etc.
2. Similarly check for ensemble model such as Random Forest, Adaboost etc.

Evaluation

1. Compare the above models based on evaluation metrics defined above.
2. Identify the best performing model and check its performance against benchmark model

8. References

1. Tatar, Alexandru, et al. "*Predicting the popularity of online articles based on user comments*" Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

2. Hensinger, Elena, Ilias Flaounas, and Nello Cristianini. "*Modelling and predicting news popularity*" Pattern Analysis and Applications 16.4 (2013): 623-635.

3. K. Fernandes, P. Vinagre and P. Cortez. "*A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*". Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.