Online News Popularity

By: Shivam Pandit, Nandini Krupa Krishnamurthy, Rohit Chundru

Introduction and Motivation

The Project aims at finding the best classification model that fits the *Online news popularity* dataset. The chosen methods for fitting the model were selected incrementally for increasing model accuracy. Implemented methods are mentioned below:

- Naive-Bayes (NB)
- K-Nearest Neighbors (KNN)
- Classification and Regression Tree (CART)
- C5.0

Data Description

The dataset – Online News Popularity – is available at https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#. It possess a set of features about articles published by *Mashable* in a period of two years.

- Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)
- Number of Datapoints: 39797

Pre-Processing and Data Cleaning

In preprocessing the data, we removed outliers in our data. Also, we removed some variables that were not important for prediction like URL and is_weekend. Some variables like data channels and weekdays were transformed to factor from integer to improve model fit. Also, shares variable needed log transformation after checking the frequency distribution graph.

Exploratory Data Analysis

We plotted frequency histograms of all the variables to analyze the variable distribution and finally used forward step wise regression model to select the best variables. Initially our dataset had 61 variables that got reduced to 30 variables using forward stepwise regression.

Data Science Model

Our goal was to find the best fit model for our dataset. We incrementally fitted models to our data set which included models like Nayive Bayes, Classification and Regression Trees, K-nearest neighbors, C5.0 to find the best data science model that accurately predicts the news popularity based on the total number of shares for the variable.

Model Evaluation

We used confusion matrix to evaluate model performance in terms of accuracy. We had fitted our model after splitting dataset in 70:30. We trained all the models on 70% data that acted as training data and tested the model accuracy on 30% data that acted as testdata. We defined articles with log(shares) larger than 7.244 (median) as popular article.

Results

- Out of the four methods implemented, C5.0 is the best to fit the model. This gives highest accuracy of 65.9%.
- The data set on a whole gives average accuracy of 61.895% which shows that the dataset is inconsistent indicating that irrelevant information has been used.
- Therefore, this data is insufficient to predict the number of shares with high levels of accuracy for a news article considering its popularity.