Final Assignment

Predicting Article Popularity in Online News Media Using Machine Learning

MBAN - Machine Learning & AI

Due on April 28, 2023 by 11:59pm

The rapid growth of online news media has transformed the way information is disseminated and consumed. With an increasing number of news articles published daily, it has become crucial for publishers and content creators to understand the factors that drive the popularity of their articles to effectively engage their audience. This project aims to leverage the Online News Popularity dataset to develop a machine learning model that predicts the popularity of news articles published on Mashable (www.mashable.com), a leading online news platform.

The dataset contains 39,644 instances and 61 features, including various content-related attributes, such as word counts, keywords, and sentiment, as well as the number of shares the articles received on social media platforms. The primary objectives of this project are to:

1. Create a regression model that predicts the number of shares an article will receive, which can serve as a continuous measure of the article's popularity.
2. Create a classification model that categorizes articles into 'popular' and 'non-popular' classes based on a predefined threshold, offering a more granular understanding of the article's popularity.

These models will enable content creators and marketers to better understand the key factors that contribute to an article's popularity and optimize their content strategies accordingly.

The project will involve the following steps:

1. Data preprocessing,
2. Model development for regression,
3. Model development for classification,
4. Model interpretation,
5. Model validation.

By the end of this project, we aim to provide reliable and interpretable machine learning models that can predict the popularity of news articles on Mashable, both as a continuous measure and as discrete categories. Additionally, we will offer actionable insights for content creators to enhance their strategies for creating and promoting engaging content.

The final assignment can be submitted as a team of maximum three people. Each section is 25 points.

Data

You can obtain the dataset from the following link: https://archive-beta.ics.uci.edu/dataset/332/online+news+popularity. Some features within the dataset may require further elaboration to ensure a clear understanding:

`kw_min_min` represents the minimum number of shares among all the articles with the least popular keyword in the metadata. In other words, it considers the keyword with the lowest popularity (in terms of the minimum number of shares) and indicates the minimum number of shares for the articles containing that keyword.

For example, if there are three keywords in an article's metadata - A, B, and C - and the minimum number of shares for articles containing each keyword are as follows:

Keyword A: 10 shares
Keyword B: 20 shares
Keyword C: 15 shares

In this case, keyword A is the least popular keyword, and the 'kw_min_min' value for this article would be 10.

`self_reference_min_shares` represents the minimum number of shares among all the articles referenced within the current article from the same source, which is Mashable in this case.

When an article links to other articles published by Mashable, it creates a self-reference. `self_reference_min_shares` captures the minimum number of social media shares of all these self-referenced articles. This can help understand the relationship between the popularity of referenced articles and the target article's popularity.

For example, if an article has three self-references with the following number of shares:

Referenced Article 1: 100 shares
Referenced Article 2: 200 shares
Referenced Article 3: 150 shares

The `self_reference_min_shares` value for this article would be 100, as it is the minimum number of shares among all referenced articles.

Latent Dirichlet Allocation (LDA) is a generative statistical model used in natural language processing and machine learning to discover hidden or latent topics in a collection of documents. It is a type of unsupervised learning method that assumes that each document in a corpus is a mixture of a small number of topics and that each word in the document is attributable to one of the document's topics.

LDA is based on the idea that words that frequently appear together in documents are likely to be related to the same topic. The algorithm represents documents as a probability distribution over topics and topics as a probability distribution over words. The main goal of LDA is to learn these probability distributions so that the model can generate similar documents when given new topic distributions.

In the context of the Online News Popularity dataset, the LDA features represent the probabilities of the article belonging to each of the five precomputed topics. These topics have been derived from the text of the articles using the LDA algorithm. The features are:

`LDA_00`: The probability of the article belonging to topic 0.
`LDA_01`: The probability of the article belonging to topic 1.
`LDA_02`: The probability of the article belonging to topic 2.
`LDA_03`: The probability of the article belonging to topic 3.
`LDA_04`: The probability of the article belonging to topic 4.

These LDA features can be used in your machine learning project to analyze the relationship between article topics and their popularity, as measured by the number of shares on social media.

The dataset does include some precomputed sentiment features that can be used for analysis. Specifically, the dataset provides the following sentiment-related features:

`global_sentiment_polarity`: the text's global sentiment polarity, which is a continuous value ranging from -1 (most negative) to 1 (most positive). It measures the overall sentiment of the text in the article.
`global_rate_positive_words`: the rate of positive words in the content. It's calculated as the number of positive words divided by the total number of words (excluding neutral words) in the text.
`global_rate_negative_words`: the rate of negative words in the content. It's calculated as the number of negative words divided by the total number of words (excluding neutral words) in the text.
`avg_positive_polarity`: the average polarity of the positive words in the text, indicating the average sentiment strength of the positive words.
`min_positive_polarity`: the minimum polarity of the positive words in the text, indicating the sentiment

strength of the least positive word.
`max_positive_polarity`: the maximum polarity of the positive words in the text, indicating the sentiment strength of the most positive word.
`avg_negative_polarity: the average polarity of the negative words in the text, indicating the average sentiment strength of the negative words.`min_negative_polarity`: the minimum polarity of the negative words in the text, indicating the sentiment strength of the least negative word.`max_negative_polarity‘: the maximum polarity of the negative words in the text, indicating the sentiment strength of the most negative word.

Steps

Before diving into the individual steps of the assignment, it is crucial to understand the overall goal and approach of the project. The primary objective of this assignment is to explore and analyze the Online News Popularity dataset, which contains various features related to news articles published on Mashable. Through a combination of descriptive (exploratory) tasks, visualization tasks, and machine learning models, you will uncover patterns and relationships between the features and the target variable, which is the popularity of the articles.

By systematically examining the dataset and its features, you will gain valuable insights into the factors that influence article popularity. These insights will not only contribute to a deeper understanding of the online news landscape but also help inform the development of more effective content strategies for publishers. The assignment is structured in a way that guides you through the data analysis process, starting with initial exploratory tasks, preprocessing, followed by visualization, and modeling, ultimately leading to the interpretation and communication of your findings.

Remember, the key to success in this assignment is a thorough exploration and understanding of the data, coupled with a methodical approach to model development and evaluation. By following the outlined steps and leveraging your knowledge of machine learning and data analysis, you will be well-equipped to tackle this challenging and exciting project.

You can also use ANN but keep in mind that ANN is not the best model for interpretation. You can still experiment with ANN in both regression and classification problems.

1. Descriptive (Exploratory) Tasks

Before diving into deeper analysis and modeling, it is essential to perform Descriptive (Exploratory) Tasks to gain a solid understanding of the dataset and its underlying structure.

1. Identifying potential data quality issues, such as missing values, outliers, and errors, which may negatively impact the performance of machine learning models.
2. Explore the relationships between pairs of features, identify potential correlations or multicollinearity that might affect the performance of your models. In the context of the Online News Popularity dataset, produce:

- **Summary statistics table:** A table showing summary statistics (mean, median, standard deviation, minimum, and maximum) for all continuous features in the dataset. This table will provide an overview of the central tendency and dispersion of the data, making it easier to understand the general characteristics of the dataset.
- **Popularity categories distribution table:** A table displaying the frequency distribution of articles categorized into ‘popular’ and ‘non-popular’ classes based on the predefined threshold. This table will help assess the balance between the two classes and provide insights into the overall popularity of articles in the dataset.
- **Data channels distribution table:** A table illustrating the frequency distribution of articles across different data channels (Lifestyle, Entertainment, Business, Social Media, Technology, and World). This table will provide insights into the prominence of various topics in the dataset and their potential impact on article popularity.

- **Weekday distribution table:** A table presenting the frequency distribution of articles published on different weekdays (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday). This table will help analyze the influence of publication day on article popularity and identify any trends or patterns related to weekdays. A table illustrating the frequency distribution of articles published on different weekdays (Monday through Sunday). This table aids in analyzing the influence of publication day on article popularity and identifying trends or patterns related to weekdays. You may perform cross-tabulation in relation to data channels and/or popularity categories.

These tables will contribute to a deeper understanding of the dataset and its underlying structure, guiding your feature selection and modeling choices. Furthermore, they offer valuable insights into factors that could influence the popularity of news articles on Mashable. When presenting these tables in your HTML or PDF files, consider using various packages that enhance the appearance and readability of tables. By ensuring a visually appealing presentation, you can effectively communicate the information and insights derived from the dataset to your audience.

3. Descriptive tasks often involve visualizations, which can help communicate your findings more effectively to your audience. Visualizations enable stakeholders to grasp complex patterns and relationships more easily, making it an essential part of any data analysis project. Use `ggplot` to plot the last 3 tables described above

2. Model development for regression

Train and evaluate various regression algorithms, such as linear regression, random forests, and gradient boosting machines, to identify the best-performing model for predicting the number of shares. Select the best tuned model that you will use in final evaluations to report the test RMSPE and its uncertainty.

3. Model development for classification

We can create a classification problem from the Online News Popularity dataset by converting the continuous `shares` feature into a categorical target variable. To do this, you can define a threshold or multiple thresholds to create categories that represent different levels of popularity.

Train and evaluate classification algorithms on the transformed dataset, such as logistic regression, decision trees, random forests, boosting methods, or neural networks.

Report the predictive performance of the best model using the test data

4. Model interpretation

Interpreting the models for both regression and classification tasks is essential to understand the key features driving article popularity and provide actionable insights for content creators and marketers. To achieve this, you can follow these steps:

- Feature importance analysis: Examine the importance of each feature in your best-performing models for both regression and classification tasks. Most machine learning algorithms, like random forests and gradient boosting machines provide methods to extract feature importance or coefficients. This information allows you to rank the features based on their contribution to the model's performance.
- Visualize feature importance: Create visualizations (e.g., bar charts) to represent the feature importance scores obtained from your models. This will help you effectively communicate the relative importance of each feature in predicting article popularity.
- Inspect the confusion table: Analyze the classification model by inspecting the confusion table.

Provide actionable insights: Based on your analysis of feature importance and interactions, offer recommendations for content creators and marketers to improve the popularity of their articles. These insights should be actionable and based on the relationships you uncovered in the dataset.

By following these steps in your model interpretation, you will be able to convey the key findings and insights effectively, supporting decision-makers in leveraging the analysis to improve their content strategies.