**ONLINE NEWS POPULARITY**

**Project Report**

**Introduction**

The aim of this project was to develop a predictive model that could estimate the popularity of an article based on its features. In order to measure popularity, the number of shares was chosen as the target variable. By predicting the number of shares, we can gain insights into the factors that contribute to an article's success and optimize future content strategies.

To accomplish this goal, I utilized a dataset provided by the UCI Machine Learning Repository, which compiled data on 39,797 articles published by Mashable. This dataset contains various characteristics of the articles, serving as valuable input variables for my predictive model.

The features used to forecast the popularity of an article included the number of words in the title and content, the rate of unique words and non-stop words in the content, the number of links and images, the average length of words in the content, the number of keywords in the metadata, the data channel (lifestyle, entertainment, business, social media, tech, or world), and the minimum, maximum, and average shares associated with the worst, best, and average keyword. Additionally, we considered the day of the week the article was published, as well as the subjectivity and sentiment polarity of the article's text and title. These features provide a comprehensive view of the article's characteristics, enabling to explore their influence on its popularity.

By developing a predictive model using these input variables, I aimed to gain insights into the key factors that drive article popularity. This knowledge can inform content creators, marketers, and publishers to optimize their strategies and increase the reach and engagement of their articles.

**Methodology**

The methodology employed in this project involved several steps to preprocess the data, extract relevant features, and apply various regression and classification models to predict the popularity of articles.

In the preprocessing phase, a total of 38 relevant features were extracted from the dataset, as described above. The schema of the data frame was defined to structure and organize the extracted features. During the preprocessing phase, several checks were also performed. Firstly, the dataset was examined for any null values, and it was determined that there were no missing values. Next, an analysis revealed that there were 1181 samples with a zero number of words in the content of the article. As these entries were considered inappropriate, they were removed from the dataset.

Categorical variables were handled by encoding them as dummy variables. Two categorical columns, "data\_channel" and "weekday," were combined to create a single column that encompassed all categories. Additionally, indexing was applied to these categorical variables.

Following preprocessing, the numeric features were scaled using standard scaling. This step ensured that all numeric variables were transformed to have a mean of 0 and a standard deviation of 1. Scaling is crucial to prevent features with larger magnitudes from dominating the models during training.

Regression models were then applied to predict the popularity of articles. The selected regression models included linear regression, generalized linear regression, decision tree regression, random forest regression, and gradient boosted tree regression. Initially, all features were used to train these models. However, in order to improve their performance, feature selection was performed using the Univariate Feature Selector. After selecting the most relevant features, the regression models were retrained. Furthermore, to explore different regularization techniques, linear regression models were also run with both ridge and lasso regularization.

In a second phase, as an attempt to improve prediction results, the target variable was classified into six categories: very poor (top 25%), poor (top 50%), average (top 75%), good (top 85%), very good (top 90%), and exceptionally good. The categorization strategy I used tries to account for the skewness of the data. This classification approach allowed the application of additional classification models. The popularity categories of the articles were predicted using multinomial logistic regression and random forest classification models.

**Results**

The results of my models are summarized in the following tables:

|  |  |  |
| --- | --- | --- |
| *Attempt 1: all features* | **RMSE** | **R-squared** |
| **Linear Regression** | 9212.01 | 0.0285 |
| **Generalized Linear Regression** | 12799.63 | 0.0156 |
| **Decision Tree Regression** | 13626.50 | -0.0129 |
| **Random Forest Regression** | 12782.60 | -0.0142 |
| **Boosted Tree Regression** | 14118.01 | -0.1718 |

|  |  |  |
| --- | --- | --- |
| *Attempt 2: features selection* | **RMSE** | **R-squared** |
| **Linear Regression** | 9221.53 | 0.0265 |
| **Generalized Linear Regression** | 12807.22 | 0.0144 |
| **Decision Tree Regression** | 13612.92 | -0.0109 |
| **Random Forest Regression** | 12571.95 | 0.0190 |
| **Boosted Tree Regression** | 13292.46 | -0.0387 |
| **Linear Regression with Ridge regularization** | 9221.53 | 0.0265 |
| **Linear Regression with Lasso regularization** | 9221.45 | 0.0265 |

|  |  |  |
| --- | --- | --- |
| *Attempt 3: classification* | **Accuracy** | **F1 score** |
| **Baseline Model** | 0.2739 | 0.1178 |
| **Multinomial Logistic Regression** | 0.3409 | 0.2437 |
| **Random Forest Regression** | 0.3499 | 0.2490 |

**Discussion**

In the first attempt, where all features were used, the performance of the regression models was limited. The linear regression model had the lowest root mean square error (RMSE) of 9212.01 and a modest R-squared value of 0.0285, indicating that only a small portion of the variance in article popularity could be explained by the chosen features. The generalized linear regression, decision tree regression, random forest regression, and boosted tree regression models also exhibited limited predictive power, with negative R-squared values in some cases.

In the second attempt, feature selection was performed to improve model performance. Despite the feature selection process, the results remained similar to the first attempt, with marginal improvements observed. The linear regression models with ridge and lasso regularization achieved nearly identical results in terms of RMSE and R-squared.

For the classification task in the third attempt, a baseline model was tested as a point of reference. The baseline model assigned the most popular class to all samples. This approach allowed for a better evaluation of the performance of the multinomial logistic regression and random forest models in comparison. The baseline model achieved an accuracy of 0.2739 and an F1 score of 0.1178, indicating poor performance. This result demonstrates the difficulty of predicting article popularity solely based on the chosen features without employing more sophisticated modeling techniques. However, the multinomial logistic regression and random forest models exhibited slight improvements compared to the baseline model. The multinomial logistic regression model achieved an accuracy of 0.3409 and an F1 score of 0.2437, demonstrating an enhancement in predicting the correct class labels. The random forest model outperformed the other models, attaining an accuracy of 0.3499 and an F1 score of 0.2490, indicating a further improvement in classification performance.

**Conclusion**

In this project, the objective was to develop a predictive model to estimate the popularity of articles based on various features. The performance of several regression and classification models was evaluated to understand the factors contributing to article success and to optimize content strategies.

The results obtained from the regression models demonstrated the challenge of accurately predicting article popularity based solely on the selected features. Despite attempts to improve performance through feature selection and regularization techniques, the models exhibited limited predictive power, and my initial goal of predicting the number of shares with a 15% error rate could not be met. The linear regression models achieved modest R-squared values, indicating that only a small portion of the variance in article popularity could be explained by the features.

For the classification task, a baseline model was established, assigning the most popular class to all samples, allowing for a comparison with more advanced models. The baseline model's poor performance highlighted the complexity of predicting article popularity based solely on the chosen features. However, the multinomial logistic regression and random forest models showed slight improvements in accuracy and F1 score compared to the baseline model. The random forest model achieved the highest accuracy, demonstrating its ability to capture some patterns and relationships within the data.

Overall, the results indicate that predicting article popularity is a challenging task, and the selected features alone may not be sufficient for accurate predictions. Further exploration and consideration of additional factors, such as user engagement metrics or social media trends, may enhance the predictive capability of the models.

This project provides valuable insights for content creators, marketers, and publishers. It emphasizes the need to consider multiple factors beyond the selected features to optimize content strategies and increase the reach and engagement of articles. By utilizing more advanced modeling techniques and incorporating additional relevant features, it is possible to develop more accurate predictive models for article popularity estimation.

Future research could focus on exploring alternative feature sets, including social media metrics, user behavior patterns, and sentiment analysis, to further improve the models' performance. Additionally, experimenting with other advanced modeling techniques and ensemble methods could provide better predictions and insights into the dynamics of article popularity.

In conclusion, while this project presents valuable insights and lays the groundwork for predicting article popularity, further refinement and exploration are necessary to develop more robust and accurate models in the rapidly evolving landscape of online content.

**References**

<https://github.com/trajanov/BigDataAnalytics/tree/master/Notebooks>

https://spark.apache.org/docs/latest/ml-features.html

<https://spark.apache.org/docs/latest/ml-features.html#feature-selectors>

<https://spark.apache.org/docs/latest/ml-classification-regression.html#regression>

<https://spark.apache.org/docs/latest/ml-classification-regression.html#classification>

<https://spark.apache.org/docs/latest/ml-pipeline.html>

https://spark.apache.org/docs/latest/sql-programming-guide.html

<https://medium.com/bazaar-tech/apache-spark-data-cleaning-using-pyspark-for-beginners-eeeced351ebf>

<https://medium.com/insiderengineering/a-deep-dive-into-sparks-univariate-feature-selector-3f8b726d7d32>

<https://runawayhorse001.github.io/LearningApacheSpark/reg.html>

<https://louisazhou.gitbook.io/notes/spark/data-cleaning-with-apache-spark>

https://hastie.su.domains/ISLR2/ISLRv2\_website.pdf

<https://www.kaggle.com/code/thehapyone/exploratory-analysis-for-online-news-popularity>