Project Report: Predicting Popularity of Online News Articles

EXECUTIVE SUMMARY: The increasing importance and availability of the internet in the 21st century have made it the primary source of new consumption via websites and social media platforms. News popularity, commonly measured by reads, likes, or shares, is a vital metric in ranking articles for stakeholders like content providers, advertisers, and news agencies. Accurately predicting popularity is thus vital for news sites, as it has financial implications in terms of advertising spend, income, and subscriptions that flows towards their websites.

This project used three machine learning models and the random forest algorithm, with an accuracy rate of 66.8% and an AOC score of 0.663, which was selected as the best to be implemented.

1.PROBLEM STATEMENT: The company currently faces the constraint of posting 72 articles per day, i.e., 12 per our 6 channels: lifestyle, business, technology, world, social media, and entertainment. It is important to select the popular articles that increase the company's ability to maximize revenue measured in terms of shares. Using the available dataset, a popularity threshold will be determined. After this, machine learning models will be built to help predict article popularity. These models will be evaluated using confusion matrix metrics and the AUC curve.

2. DATA DESCRIPTION: The dataset consists of 39,643 and 61 attributes, which include 1 target attribute (number of shares) and 2 non-predictive features (article URL and days between publication and dataset acquisition). It had no missing or duplicate values, and the categorical features like article channel and day of publishing were preprocessed into numeric features. A new column, named "popular," was generated by utilizing the data from the "shares" column. The criteria for judging article popularity were based on the median (1400) instead of the mean (3395) of the shares column due to the presence of outliers. Articles with fewer than 1400 shares were categorized as unpopular, while articles with 1400 shares or more were considered popular. The values in the popular column were encoded as 0 and 1, respectively.

After this, columns with a 'num' prefix were visualized using histograms to view their counts, while a box plot was used to show the relationship between popular and columns like article type and day of week.

Lastly, redundant attributes that fall into the following categories: non-predictive (2), repetitive (7 including days of the week only), and multicollinear (12) were dropped to reduce model complexity and the risk of overfitting.

3. MODEL ANALYSIS: Given the nature of the problem of determining if news articles will be popular or not, three classification models were used to predict the dependent variable. The data was split into a training set of 75% and a test set of 25%. The dependent variable 'popular' was also separated from the 37 independent variables that made it to the modelling stage.

Logistic regression is used to predict binary outcomes, and its output lies between 0 and 1, with a probability greater than 0.5 indicating event occurrence (in this case, popularity). This model had an accuracy of 64.4%, and it did better than the Naïve Classifier with an accuracy of 53.6%. The naïve classifier assigns all data points to one class and is used to test model performance.

Decision trees visually represent decision-making processes using an upside-down tree structure. It automates complex decision-making by partitioning data into subsets based on specific criteria. Three partitions of this model (cp = 0.01, cp = 0.001, and cp = 0.00001) were run, and the second partition (cp = 0.001) performed best with an accuracy score of 64.5%.

Random Forest enhances decision tree performance by combining multiple trees trained on different data and features to improve accuracy and reduce overfitting. Two iterations of this model were used, with ntree (number of trees) = 500 coming out on top with an accuracy of 66.8%.

Lastly, the random forest model was also used to calculate feature importance scores which measures the contribution of each feature to the overall predictive ability of the model.

4. MODEL EVALUATION: Models were evaluated using two core classification metrics: accuracy and AUC score. Accuracy in this context measures the correct predictions made by each model in relation to the total number of predictions made. AUC measures the ability of the model to distinguish between two classes, typically the positive class (popular) and the negative class (unpopular). Random Forest scores highest on both metrics and has been selected as the model to be implemented.

Model	**Accuracy**	**Testing Error**	**AUC Score**
Logistic Regression	64.4%	35.6%	0.646
Decision Trees	64.5%	35.5%	0.647
Random Forest	**66.8%**	**33.2%**	**0.663**

5. RECOMMENDATIONS AND CONCLUSION: To summarize the project, a dataset of 39,463 articles with 61 features was used to predict the popularity of news articles. This data was analyzed, visualized, and preprocessed before feeding into three machine learning models. The most optimal result was random forest, which had a prediction accuracy of 66.8% and an AUC score of 0.663. The top 3 features that contributed to accuracy of the model where: Avg. key words (max. shares), self-reference avg. shares and weekend posts. Articles with these attributes should be prioritized.

To further improve predictive performance, the following approaches can be taken: gathering more data, incorporating additional features into the original dataset, and lastly, implementing advanced techniques such as natural language processing and large language models to predict article popularity.