Predicting Social Share Counts for Online News Articles

A MACHINE LEARNING APPROACH

Arunsingh Kopalakrishnaswami
Assistant Manager - Supply Chain Engineering

Executive Summary

A Comprehensive Overview of Predictive Modeling for Article Popularity

Objective:

- Predict the number of times an online news article will be shared before publication.

Dataset:

- *Online News Popularity Dataset*

Approach:

- Analyze factors contributing to an article's popularity.
- Develop predictive models to forecast article shares.

Goal:

- Provide insights into the key elements driving online engagement.
- Deliver a predictive framework for future articles' success.

Rationale

Why Should Anyone Care About This Question?

- **Content creators** seek to maximize the reach and impact of their articles.
- **Marketers** benefit from knowing which factors drive online engagement to optimize digital campaigns.
- **News agencies** can strategically plan content based on predicted popularity, enhancing user engagement.
- **Publishers** can improve advertisement strategies by targeting articles with higher predicted shares.
- **Researchers** gain insights into how information dissemination works in the digital age.

Research Question

What Are We Trying to Answer?

1. What factors influence the number of times an online news article is shared?
2. Can we develop a model to accurately predict the popularity of an article before it is published?
3. How can these insights help content creators, marketers, and publishers make data-driven decisions?

The goal is to identify key drivers of article popularity and forecast the number of shares effectively.

Data Sources

Leveraging the Online News Popularity Dataset for Predictive Analysis

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).

Dataset Overview:

- **Title:** Online News Popularity
- **Source:** Mashable (www.mashable.com)
- **Type:** Multivariate
- **Instances:** 39,644
- **Features:** 61
- **Missing Values:** No
- **Reference URL:** `https://archive.ics.uci.edu/dataset/332/online+news+popularity`

Methodology

Applying Machine Learning Techniques for Accurate Share Prediction

Data Preprocessing:

- Handle missing values and outliers. Feature scaling and encoding categorical variables.

Exploratory Data Analysis (EDA):

- Identify trends and correlations between features, then visualize the relationship between key variables and article shares.
- Univariate Analysis: Analyze individual variables using summary statistics and visualizations (e.g., histograms, box plots).
- Bivariate Analysis: Analyze the relationship between two variables with scatter plots, correlation coefficients, and cross-tabulations.
- Multivariate Analysis: Investigate interactions between multiple variables using pair plots and correlation matrices.

Methodology

Applying Machine Learning Techniques for Accurate Share Prediction

Modeling Approaches:

- Build and compare models:
  - Dummy Regression (baseline)
  - Linear Regression
  - Decision Tree Regression
  - Support Vector Machine (SVM) Regression
  - Random Forest
  - XGBoost
- Use cross-validation for model evaluation then fine-tune models using hyperparameter optimization.

Model Comparison and Selection:

- Compare models and select the best model for predicting article shares.

Results

Key Findings from Model Comparisons and Predictive Performance

Model	MAE	MSE	RMSE	R-squared	MAPE
Dummy	2.899458e+03	5.109779e+07	7.148272e+03	-0.000727	2.323477e+02
Linear Regression	1.224694e-11	2.505465e-22	1.582866e-11	1.000000	1.523497e-12
SVM Regressor	2.115428e+03	5.329547e+07	7.300375e+03	-0.043767	8.101401e+01
Decision Tree	1.207952e+01	7.426002e+04	2.725069e+02	0.998546	1.219213e-01
RandomForest	7.649595e+00	3.823289e+04	1.955323e+02	0.999251	2.607302e-01
XGBoost	2.756918e+02	4.621107e+07	6.797872e+03	0.094977	2.917349e+00

- **Best Model**: RandomForest exhibits the most accurate predictions with the lowest errors across all metrics, followed closely by Decision Tree.
- **Poor Performers**: The SVM Regressor and Dummy models perform poorly, with high errors and negative R-squared values, indicating poor fit to the data.
- **Potential Improvement Areas**: XGBoost and SVM could benefit from hyperparameter tuning to improve their predictive power.

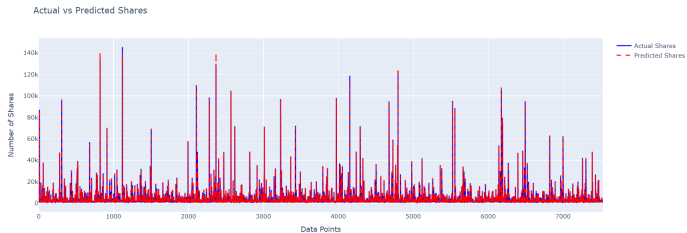
Results

Key Findings from Model Comparisons and Predictive Performance

Forecasted Details:

	Actual Shares	Predicted Shares
32790	1200	1200.0
12867	711	711.0
39154	843	843.0
23118	3400	3400.0
5054	3400	3400.0

RandomForest Regressor Test Data Metrics:
MAE: 7.6496
MSE: 38232.8934
RMSE: 195.5323
R-squared: 0.9993
MAPE: 0.2607



After BayesSearchCV Hyperparameter Tuning:

Best parameters found by BayesSearchCV: OrderedDict({'bootstrap': True, 'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 100})
RandomForest Regressor (Best) - MSE: 24927.944849637468
RandomForest Regressor (Best) - R^2: 0.9995117975978828

Conclusion:

The Random Forest Regressor, after tuning and cross-validation, shows promising performance in predicting the number of shares. The tuned model provides a high R-squared value, indicating its effectiveness in explaining the variance in the target variable. The MSE is also relatively low, suggesting that the model performs well in terms of accuracy.

Next Steps

Recommendations for Future Enhancements and Model Optimization

- **Hyperparameter Tuning**: Continue to refine models like XGBoost and SVM for better performance.
- **Feature Engineering**: Explore additional features or transformations to improve model accuracy.
- **Deployment**: Implement the RandomForest model in a production environment to predict shares in real-time.
- **Further Analysis**: Investigate specific features' impact on share counts for deeper insights.

Any Doubts or Clarifications?

Please feel free to reach out to me at:
arunsingh.kks@gmail.com

Thank you for your attention!