**Predicting Online News Popularity with Machine Learning Report**

Group 1: Anqi Gu, Evelyn Zhou, Han Zhang, Hongyu Liao, Yiwei Li

MASTER OF SCIENCE IN DATA SCIENCE

MSDS 422 – Practical Machine Learning

March 11, 2025

### **Chapter 1: Executive Summary**

With digital media growing rapidly, understanding what makes online news articles popular is essential for media companies, advertisers, and content creators. This project applies machine learning to predict news article popularity using the Online News Popularity dataset, which includes 39,797 articles from Mashable. The dataset has 58 numerical features covering textual, structural, and contextual aspects (Fernandes, Vinagre, and Cortez 2015).

Following the CRISP-DM methodology, the project involves data preprocessing, exploratory data analysis (EDA), feature engineering, and predictive modeling. Articles are categorized into two groups based on their share counts. Machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, and a Neural Network model (MLPClassifier), are trained and evaluated to determine the most effective approach. The study also investigates the most influential features contributing to shares, examining linguistic, visual, and structural characteristics that significantly influence model performance.

The results provide practical insights for media professionals looking to optimize content strategies through data-driven decision-making. Additionally, considerations for real-time deployment and automation are discussed, demonstrating how predictive analytics can enhance digital content optimization(Arora and Sharma 2023).

### **Chapter 2: Problem Statement/Research Objectives**

As online content consumption increases, media organizations face challenges in understanding what drives news article popularity. Traditional approaches rely on manual curation and basic heuristics, which do not scale effectively in a high-volume environment. This research aims to develop a machine learning-based approach to predict article popularity using textual and contextual features.

In this project, we examine how different factors contribute to article shares, analyzing the impact of textual content, structural elements, and engagement indicators. It also compares the performance of multiple machine learning models to determine which method yields the most accurate predictions. Additionally, the project explores how preprocessing techniques, such as feature scaling and transformation, affect model performance. Finally, the study considers practical aspects of deploying an automated system for real-time popularity prediction, discussing potential business applications and technical constraints.

By addressing these areas, we seek to create a scalable, data-driven solution for media organizations looking to optimize their content strategies.

### **Chapter 3: Literature Review**

Research on online news popularity prediction has explored various factors influencing content virality. Prior studies emphasize that textual, structural, and contextual features play a key role in determining how widely an article is shared. Tsai and Wu (2022) used Random Forest, LightGBM, and XGBoost to analyze the Mashable dataset, highlighting how elements like article topics, title length, multimedia presence, and sentiment polarity affect engagement. Their study achieved an 88% accuracy rate, demonstrating that feature engineering improves prediction performance.

Shang et al. (2022) expanded on this by incorporating social influence and homophily effects into prediction models. They used variational graph autoencoders (VGAE) and attention-based graph neural networks to analyze interactions between users and content. Their findings suggest that early adopters play a crucial role in content diffusion, and audience segmentation can refine predictive models.

Rajagopal et al. (2022) took a different approach, incorporating psycholinguistic features such as readability scores, emotional tone, and psychological markers. Their study found that articles with clearer language and emotionally engaging content were more likely to be shared, reinforcing the importance of linguistic analysis in popularity prediction.

Other research focuses on integrating multimodal data for better accuracy. Arora et al. (2023) proposed a model that combines text and image features using ensemble learning. Their study found that visual elements significantly impact engagement, often as much as textual components. Meanwhile, Fernandes et al. (2015) introduced an Intelligent Decision Support System (IDSS) that integrates predictive modeling with content optimization strategies. Their system suggests modifications to headlines and keywords to increase an article’s reach, offering a more actionable approach to improving content performance.

Despite progress in this field, several challenges remain. Data imbalance, evolving user behavior, and shifting content trends make long-term accuracy difficult. Many studies focus on short-term prediction without considering how engagement patterns change over time. This project builds on previous research by integrating multiple machine learning approaches while emphasizing feature selection and model interpretability. By combining traditional techniques with deep learning, the study aims to improve predictive accuracy and provide a practical framework for media organizations.

### **Chapter 4: Data Preparation**

Before beginning our exploratory data analysis, we implemented several data preprocessing steps to ensure data quality and optimization for modeling.

The UCI Online News Popularity dataset was complete with no missing values, allowing us to focus on optimizing features rather than addressing missingness.

To enrich the dataset with writer-specific information, we implemented web scraping techniques to extract author names and follower counts, adding valuable dimensions that reflect writing style and quality. These additional features provided insights beyond the physical structure of articles, capturing author influence and audience engagement patterns that weren't present in the original dataset.

For the target variable (shares), we implemented a median-based binary classification approach. Through data analysis, we identified 1400 shares as the precise median of the dataset, dividing articles into two categories: low-sharing (below 1400) and high-sharing (greater than or equal to 1400). This approach created a relatively balanced classification problem with 46.64% of articles in the low-sharing category and 53.36% in the high-sharing category. Choosing the median as the threshold proved more robust than using the mean (3395.38), as the latter is susceptible to extreme values.

### **Chapter 5: Exploratory Data Analysis**

This dataset consists of 39,644 rows and 65 columns, with a memory usage of 19.66 MB. The data type distribution is as follows: 34 columns are float64, 25 columns are int64, and 6 columns are object type, indicating that the dataset is primarily numerical, with a small number of categorical features.



Figure 5.1. Missing values and followers analysis



Figure 5.2. Social Media Dataset Statistics: Followers, Channels, Day Types, and Author Levels

This dataset has no missing values, indicating high data quality. It includes six categorical (object) variables, namely shares, followers, followers\_original, channel, day\_type, and author\_level, which are mainly used for classification.

* Shares (Share Level): Only has two categories: High and Low, with 21,154 records labeled as High and 18,490 as Low, showing a relatively balanced distribution.
* Followers (Follower Type): Contains seven different categories, with the most common being Low (15,234 records), followed by Unknown (8,643 records). Other categories such as Medium, Reprinted, and Extremely Low also appear frequently.
* Followers\_original (Original Followers Count): Has 151 unique values, with Null (9,297 records) and Reprinted (3,747 records) being the most frequent, indicating that many records lack original follower data.
* Channel (Article Channel): Includes seven distinct categories, with the most common being world (8,427 records), tech (7,346 records), and entertainment (7,057 records), suggesting a diverse range of content.
* Day\_type (Publishing Day Type): Divided into weekday and weekend, with weekday records (34,454) being significantly higher than weekend records (5,190), indicating that most data comes from weekdays. Calculating the averages reveals a high level of content activity on weekdays, with average daily records on weekdays being more than 2.5 times higher than on weekends.
* Author\_level (Author Level): Consists of Medium (22,089 records), High (17,360 records), and Low (195 records). Most authors fall into the Medium or High category, while Low-level authors are rare.
* Overall, this dataset mainly consists of numerical and categorical variables, with no missing values. The distributions are relatively balanced, and key variables such as shares, followers, and author\_level may play an important role in future analyses.

Next, we will analyze the distribution and characteristics of the target variable “Shares”. We plot the pie chart and bar chart to show the distribution.

The target variable shares (article shares) is categorized into High and Low:

* High: 53.36% (21,154 articles)
* Low: 46.64% (18,490 articles)

The number of highly shared articles is slightly higher than those with low shares, but the difference is not substantial.

The Pie Chart below displays the proportion of articles in each category, indicating that achieving high virality is challenging. While the High category is slightly larger, a significant portion of articles still falls under Low shares. Understanding what drives virality is crucial.

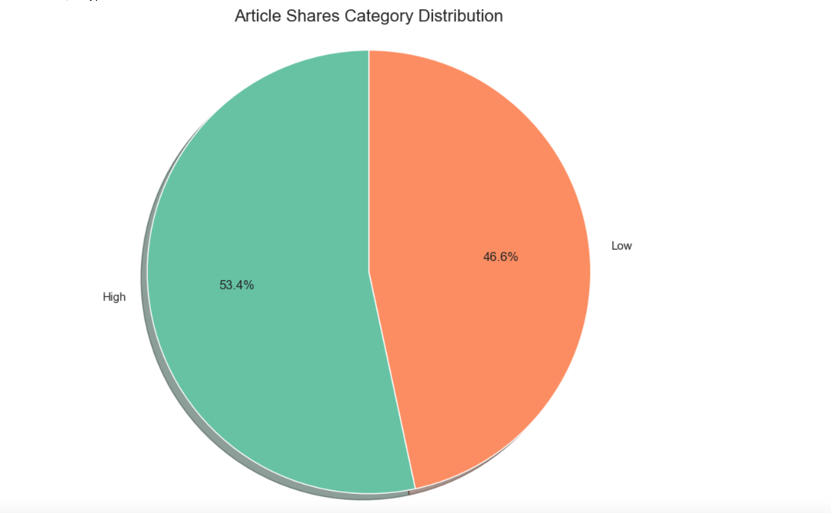


Figure 5.3.Pie chart of article shares category distribution

The Bar Chart below quantifies the number of articles in each share category, reinforcing the observation that most articles receive lower engagement, with only a small fraction achieving viral success.

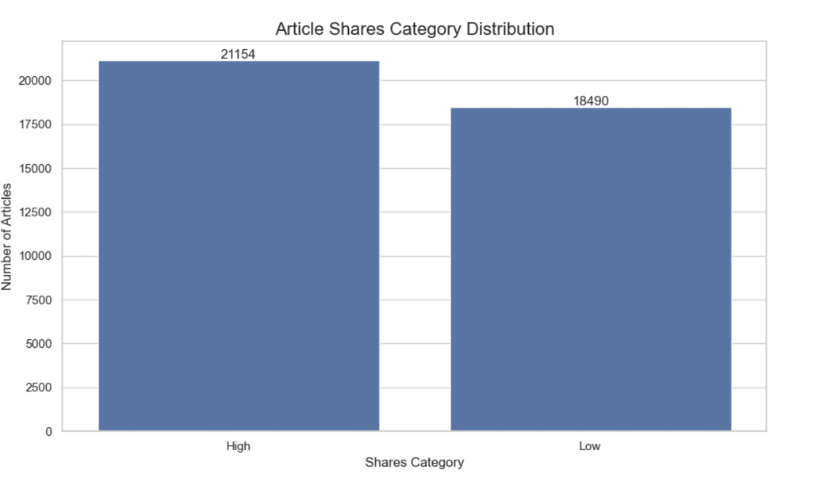
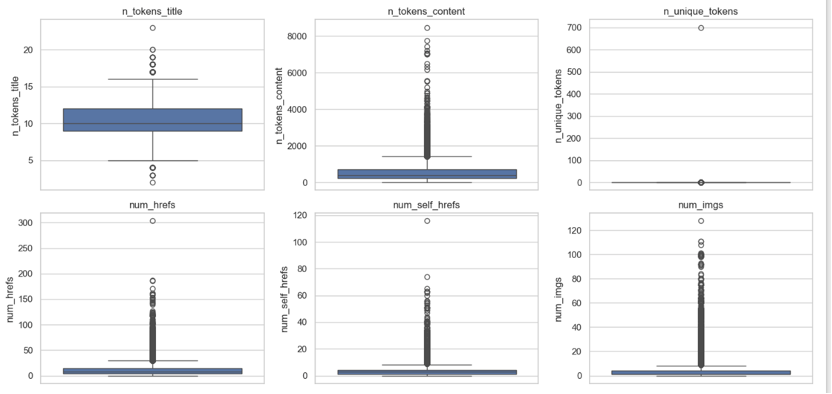


Figure 5.4.bar chart of article shares category distribution

In conclusion, most articles have lower share counts, with only a small fraction widely shared. Analyzing factors such as content type, posting time, and author influence can help improve content strategies and maximize article reach.

The next section analyzes key content features of articles, including title length, body length, hyperlinks, images, videos, and keywords. The average title length is 10.4 words, while body length varies significantly, averaging 546.5 words, with some exceeding 8,000 words. Articles contain an average of 10.88 hyperlinks and 3.29 self-referencing links, often used for SEO optimization. Multimedia usage varies, with some articles having no images or videos, while others contain up to 128 images or 91 videos. Keyword count averages 7.22, influencing SEO strategies.

Overall, article content features show significant variation, particularly in body length, multimedia usage, and hyperlinks. Keywords and self-referencing links may impact search visibility and shareability. Further analysis can explore which features influence article shares, optimizing content strategy for better engagement.



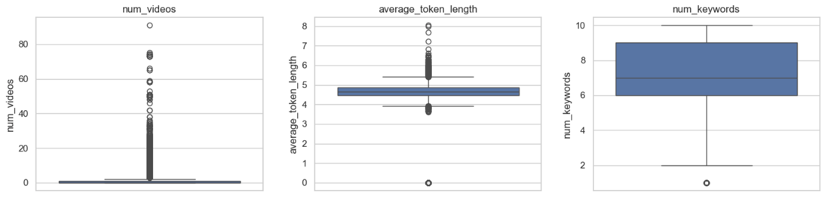


Figure 5.5.Content Metrics Distribution: Text, Links, Media, and Keywords

The boxplots highlight highly skewed distributions in various article features, with numerous outliers indicating that some articles contain exceptionally high values. Title word count remains relatively stable, typically between 9 and 12 words, while body word count varies significantly, with some articles being long-form reports exceeding the norm.

Features like hyperlinks, self-referencing links, images, and videos show heavy-tailed distributions, meaning most articles contain only a few, but a small number have exceptionally high counts, likely for SEO or multimedia engagement purposes. In contrast, keyword usage is more balanced, usually ranging between 6 and 9 per article, suggesting a consistent SEO strategy across most content.

Overall, the dataset exhibits substantial variation in article characteristics, requiring data preprocessing, such as normalization, to mitigate the impact of extreme values and ensure robust predictive modeling.

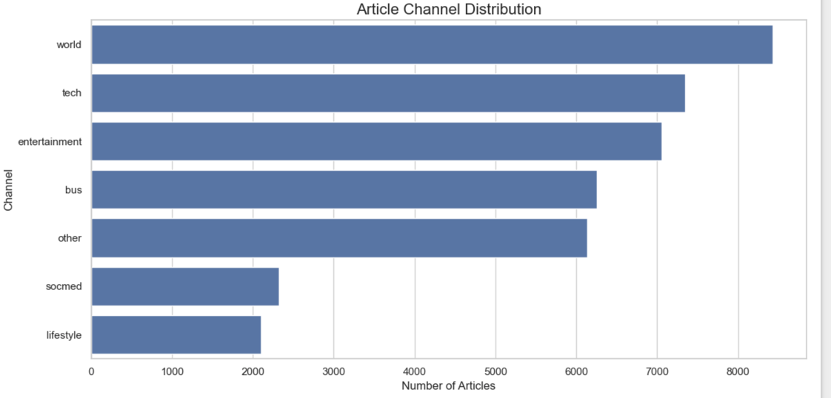


Figure 5.6. Article Chanel Distribution

The analysis reveals that world news, technology, and entertainment are the most frequently covered topics, with significantly fewer articles in social media and lifestyle categories. This suggests that both audience preferences and platform strategies prioritize global affairs, technological advancements, and entertainment content, while social media trends and lifestyle topics receive less attention.

This distribution highlights content priorities in media publishing, where certain topics are emphasized to maximize engagement. Understanding these trends can help content creators refine their publishing strategies to align with audience demand and optimize content reach.

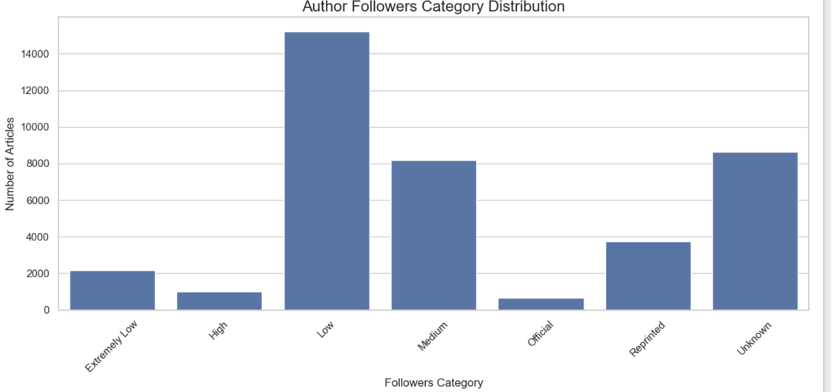


Figure 5.7. Author Followers Categoryl Distribution

The analysis reveals that the majority of articles are written by authors with low or unknown follower counts, with high-follower authors contributing significantly less. This suggests that less influential authors actively publish more frequently, likely to gain visibility and attract readership, while highly followed authors may prefer other content distribution channels. Additionally, a substantial portion of articles is reprinted from external sources, indicating that a notable share of the platform's content is not original but republished.

This distribution highlights different levels of author influence on the platform, where low-follower authors dominate content creation, while high-follower or official authors contribute sparingly. Further analysis could explore whether an author's follower count correlates with article shares, providing insights into how influence affects content virality and engagement.

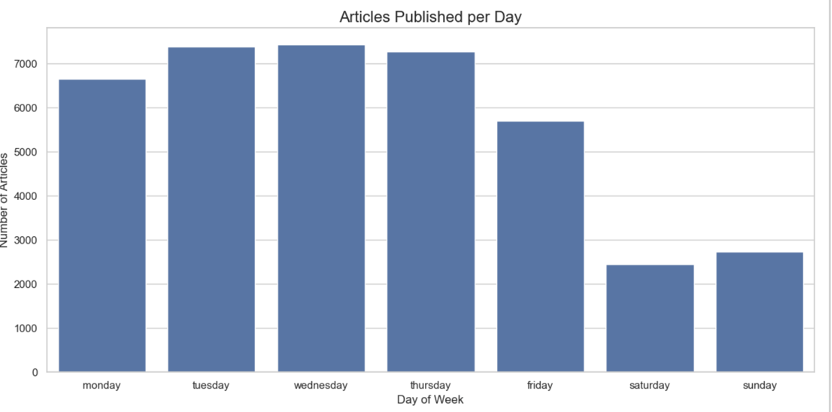


Figure 5.8. Articles Published per Day

The analysis shows that article publication is highest on weekdays, particularly on Tuesday and Wednesday, while weekend publications drop significantly. This suggests that content creators prioritize weekdays when user engagement is higher, whereas weekends see fewer articles due to lower reader activity or platform strategies. Further analysis could examine how publication timing affects reach and engagement, helping optimize content scheduling.

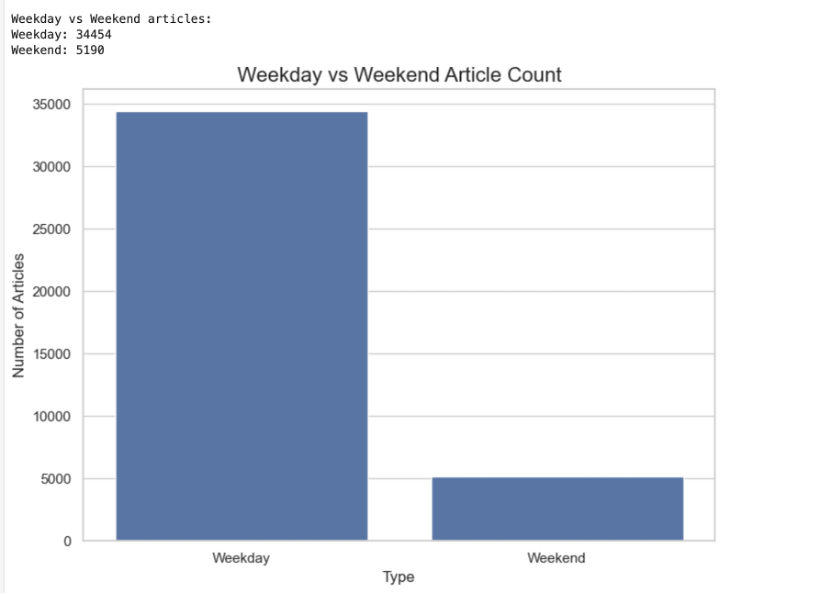


Figure 5.9. Weekday vs Weekend Article Count

The analysis reveals a significant difference in article publication between weekdays and weekends, with 34,454 articles published on weekdays compared to just 5,190 on weekends. The sharp decline in weekend publications suggests that media platforms prioritize content during weekdays, likely due to higher audience engagement and newsroom activity during these periods.

The reduced publication rate on weekends may be due to lower reader activity, leading to a strategic shift in content scheduling. This trend highlights the importance of timing in content distribution, as publishing during peak engagement periods can maximize reach and interaction. Further analysis could examine how publication timing affects article shares, helping refine strategies to enhance visibility and audience engagement.

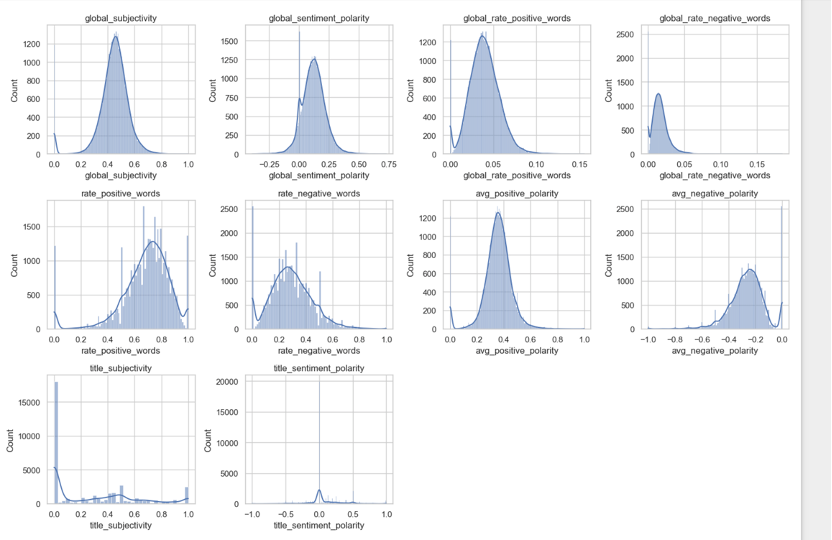
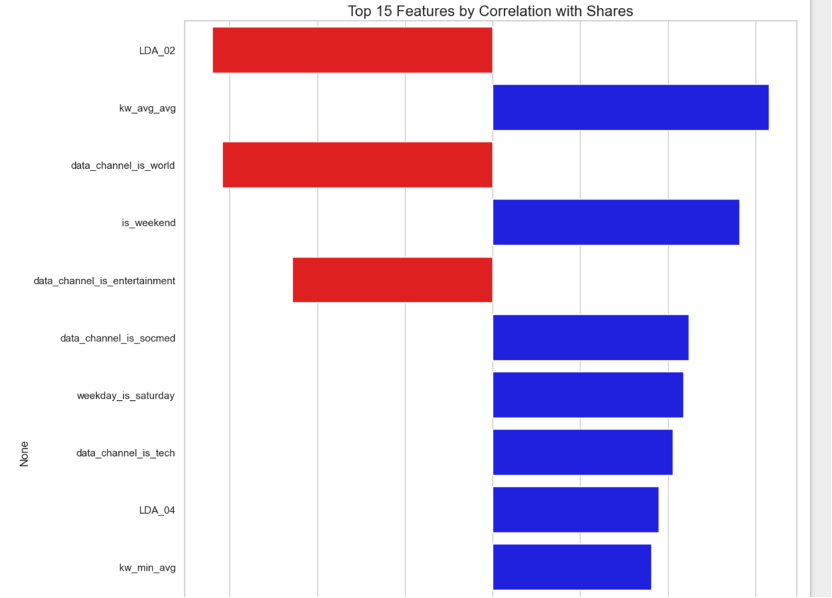


Figure 5.10.Sentiment and Subjectivity Distribution Analysis

The analysis shows that articles are generally moderately objective with a slight positive sentiment bias. The global subjectivity averages 0.44, indicating a balance between objectivity and subjectivity, while global sentiment polarity (0.12) suggests a mild positive tone. Articles contain more positive words (0.62) than negative words (0.29), and most titles remain neutral, though some exhibit extreme sentiment.

While most articles maintain a balanced tone, some show strong negativity or high subjectivity. Further analysis could explore how sentiment influences article shareability, helping refine content strategies to enhance engagement.



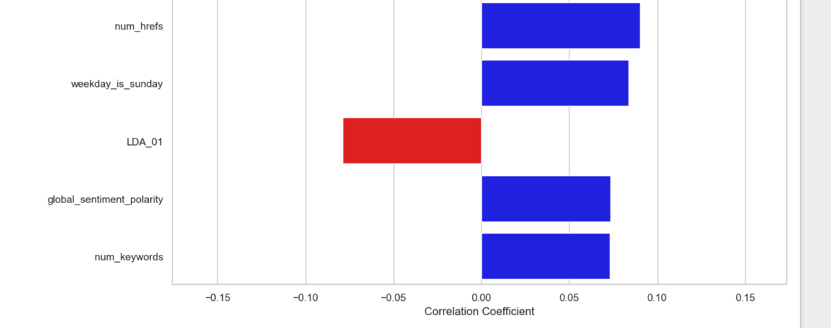


Figure 5.11.Top 15 Features by Correlation with Shares

The analysis highlights key factors influencing article shares, revealing that keywords, weekend publication, social media presence, and hyperlinks are positively correlated with higher engagement. This suggests that well-optimized keywords, strategic weekend posting, and increased hyperlink usage can enhance article visibility and shareability.

Conversely, certain LDA topics and content categories like "World" and "Entertainment" show negative correlations, indicating that despite their high publication volume, they receive fewer shares, possibly due to lower audience engagement. This insight suggests that some widely covered topics may not necessarily drive high interaction.

Overall, content optimization should focus on effective keyword strategies, leveraging social media, and refining publication timing to maximize audience reach. Further analysis could explore how these features interact with other engagement metrics, helping to refine content strategies for better performance.

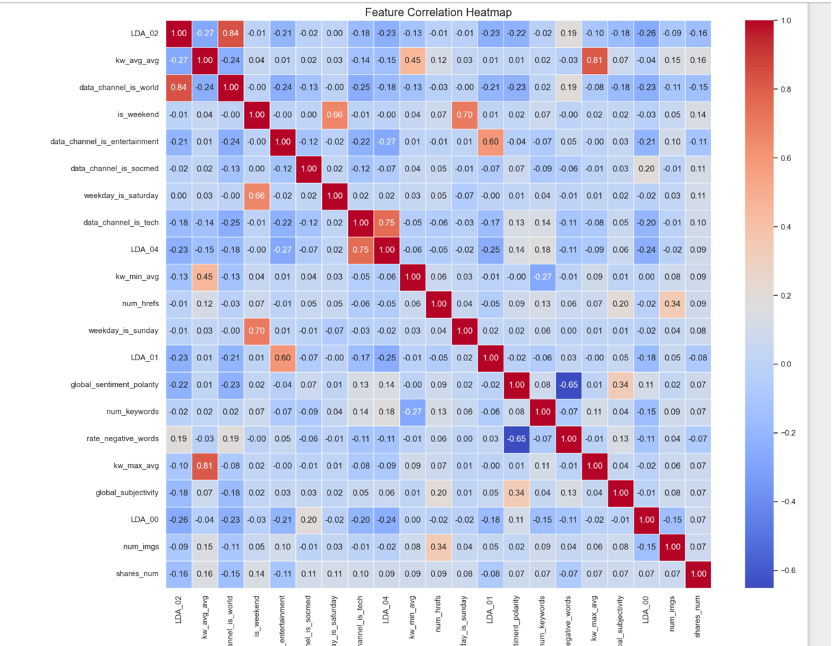


Figure 5.12.Feature Correlation Heatmap

The heatmap confirms previous findings, showing that keyword performance (kw\_avg\_avg), hyperlinks (num\_hrefs), and weekend publication (is\_weekend) positively correlate with article shares. In contrast, LDA\_02 and world news articles (data\_channel\_is\_world) show negative correlations. This suggests that optimized keywords, links, and timing can boost engagement, while certain topics may receive fewer shares.

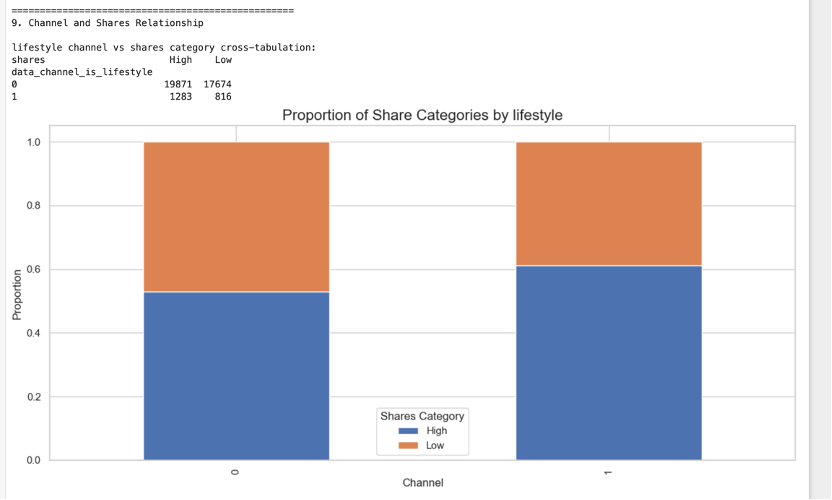


Figure 5.13.Proportion of Share Categories by lifestyle

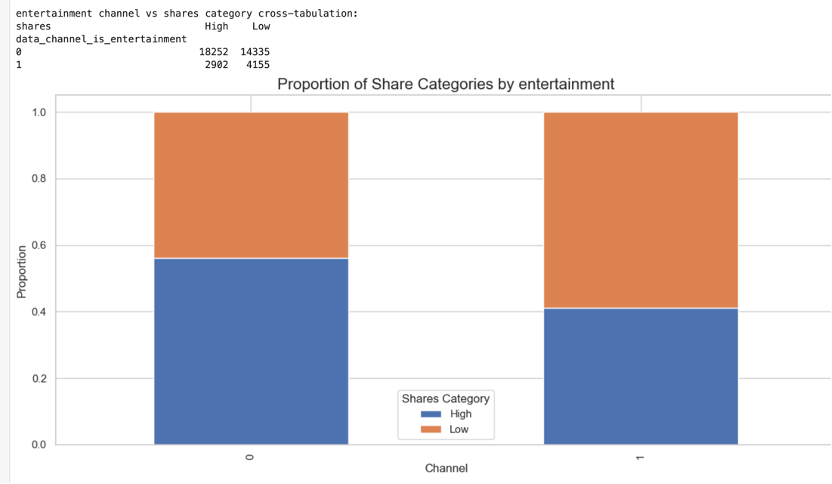


Figure 5.14.Proportion of Share Categories by entertainment

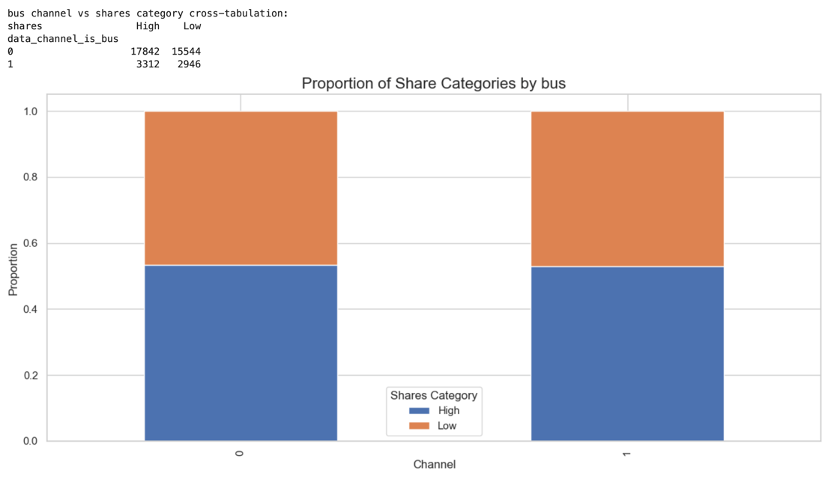


Figure 5.15.Proportion of Share Categories by bus

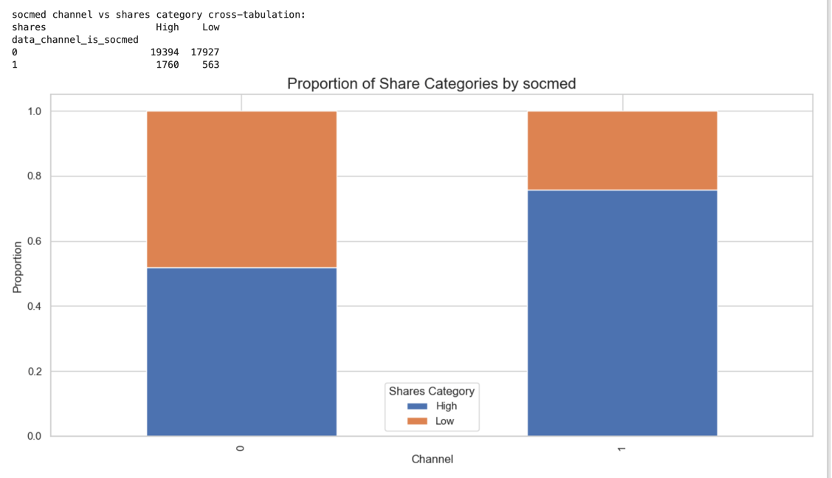


Figure 5.16.Proportion of Share Categories by socmed

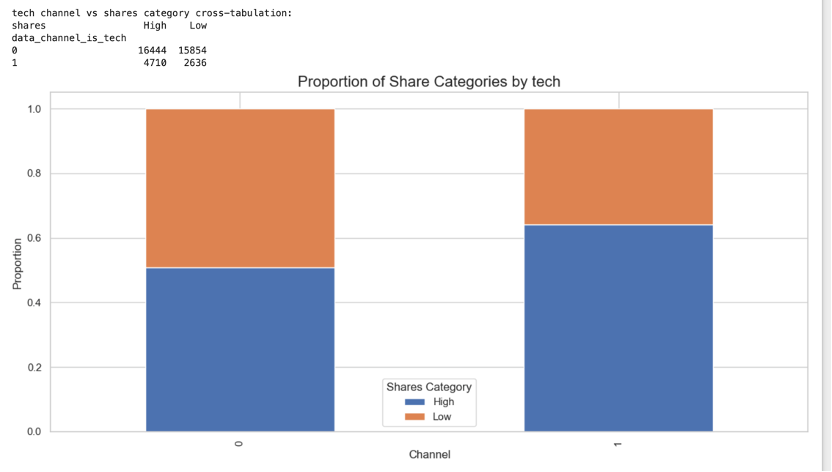


Figure 5.17.Proportion of Share Categories by tech

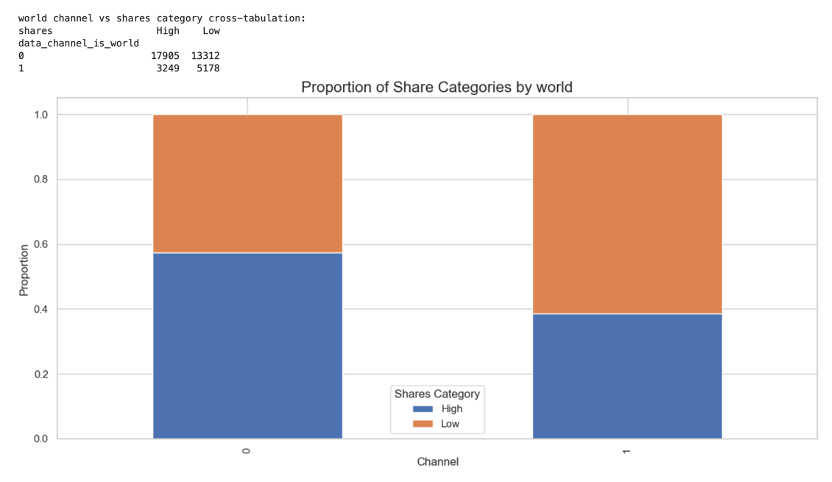


Figure 5.18.Proportion of Share Categories by world

The analysis examines the relationship between different content channels and article shares. The results show that across all categories, there is a relatively balanced distribution of high and low share articles, with some variations. The lifestyle channel has a lower total number of articles, and a slightly lower proportion of high-share articles compared to other categories. Entertainment and business channels have a more even distribution between high and low shares, indicating that audience engagement in these categories is relatively stable. The social media (socmed) channel stands out, with a higher proportion of high-share articles, suggesting that social media-related content tends to receive more engagement. The tech and world channels also display a balanced share distribution, though the world news category has a slightly higher proportion of low-share articles.

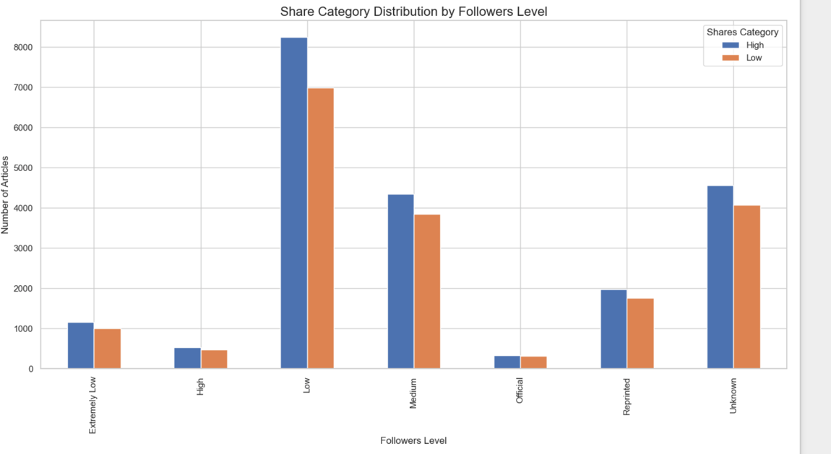


Figure 5.19.Share Category Distribution by FFollowers Level

The analysis explores the relationship between follower levels and article shares. The results indicate that articles from low-follower authors have the highest publication volume and also receive the most shares, followed by those from medium-follower authors. Interestingly, articles from high-follower authors have the lowest count, suggesting that highly followed authors contribute less frequently. Reprinted and unknown-author articles also show notable engagement, with high-share articles slightly outnumbering low-share ones.

Overall, the data suggests that low and medium-follower authors drive the majority of content and engagement, while high-follower authors contribute less frequently. This could indicate that active, lower-follower authors work harder to gain visibility, whereas highly followed authors may rely on other distribution methods for engagement. Further analysis could examine whether follower levels directly impact article virality.

### **Chapter 6: Feature Engineering**

Based on insights gained from our exploratory data analysis, we implemented several feature engineering techniques to enhance model performance.

To address the highly skewed distributions of several features identified during EDA, we implemented a two-step transformation process. First, we applied Box-Cox transformations to normalize the distributions of positively skewed variables. Second, we implemented capping techniques to handle extreme values, replacing outliers above the 99th percentile with the 99th percentile value. This combined approach effectively normalized the distributions while preserving the overall data structure and relationships.

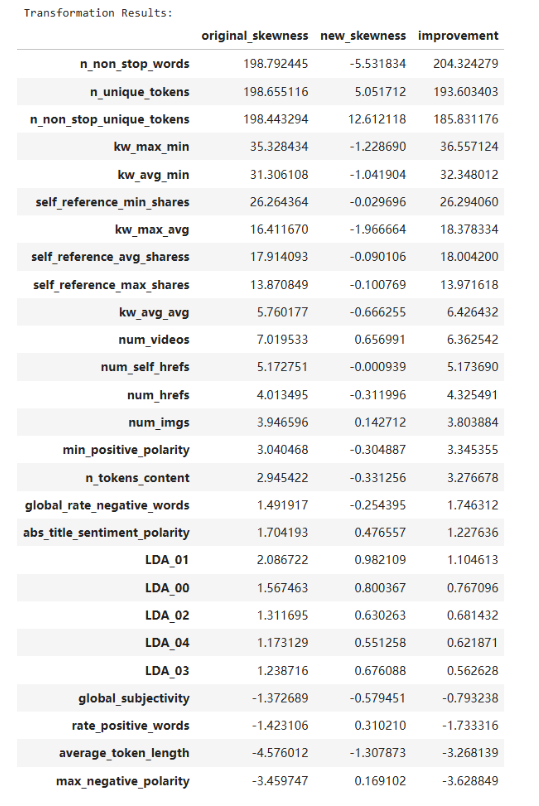


Figure 6.1. Skewness Transformation Result

We also applied min-max scaling to ensure that all numerical variables were normalized to a common range. These transformations improved model convergence, particularly for gradient-based algorithms.

Although we initially experimented with feature selection techniques to reduce dimensionality by removing features with over 95% correlation, we observed performance degradation in the resulting models. Therefore, we made the data-driven decision to retain all features in the final modeling approach, allowing the algorithms to leverage the full feature space. This approach preserved potentially subtle but important interactions between features that might have been lost through aggressive feature reduction.

These preparatory steps ensured that the dataset was optimized for machine learning models, balancing interpretability with predictive accuracy. Converting the shares variable into a binary classification problem simplified the prediction task, allowing models to focus on distinguishing potentially successful content from less likely successful content, providing clear decision support for content creation and promotion strategies.

### **Chapter 7: Methodology and various tools used in the process**

This project was conducted using Anaconda, a Python virtual environment manager that is frequently used in data science and machine learning workflows. The model training and analysis were performed in a Jupyter Notebook, leveraging the flexibility and interactivity of Python-based environments. For hardware, all computations were executed on the local CPU. All the codes and scripts have been uploaded to Github repository.([Github Link](https://github.com/hongyu-liao/MSDS-422-Final-Project-Group-1))

We employed six models for our binary classification task: Logistic Regression, Random Forest, XGBoost, LightGBM, Gradient Boosting, and Neural Networks (MLPClassifier from Scikit-Learn). Random Forest was selected for its ensemble approach that mitigates overfitting. Gradient Boosting sequentially corrects prediction errors. XGBoost and LightGBM, the two most commonly used models, were set up to target binary classification objectives, offering strong predictive power for structured data. MLP is not the model we were focusing on, just for comparison.

Our study found that different models predict differently for different classifications. In order to synthesize the performance of different models, we will use the model stacking technique. We first build models using LightGBM, XGBoost and Logistic Regression fitting data called Level 0 models. Then their predictions are fed into the Gradient Boosting model as new training data, called Level 1 model. The final result obtained combines the performance of these models.

Performance was evaluated using cross-validation (K-Fold), with accuracy as the primary metric for our binary classification task. Besides, we plot confusion matrices for every model. The accuracy metric represents the percentage of articles our model correctly identifies as either highly shareable or not, providing a clear measure of prediction reliability for making data-driven decisions. Confusion matrices can provide detailed analysis of true and false classifications. For hypertuning, we used GridSearchCV and RandomizedSearchCV for parameter search to try to find the best combination of parameters based on the demand of computational resources.

Our optimization workflow employs a two-stage parameter tuning method to balance efficiency and performance. The process begins with a coarse search using a smaller sample (5,000 observations) and broader parameter ranges with 3-fold cross-validation. This initial stage identifies promising regions in the parameter space. Then we refine these parameters in the second stage using a larger sample (10,000 observations) and 5-fold CV for more robust validation. After determining optimal hyperparameters, we train the final model on the complete dataset and evaluate performance using accuracy and F1-score metrics.

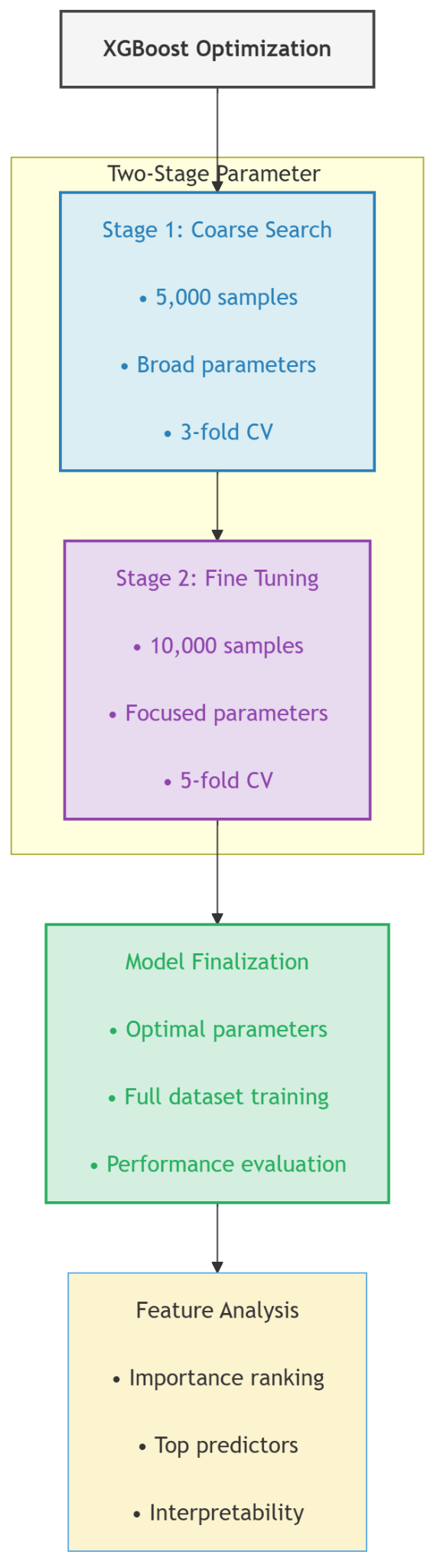


Figure 7.1. Workflow of Optimization

### **Chapter 8: Findings and Conclusions**

The machine learning models demonstrated varying levels of predictive accuracy in classifying article shareability. We found that Optimized XGBoost achieved an accuracy of 68.12%, followed closely by the Stacking model at 67.69% and LightGBM at 67.39%.

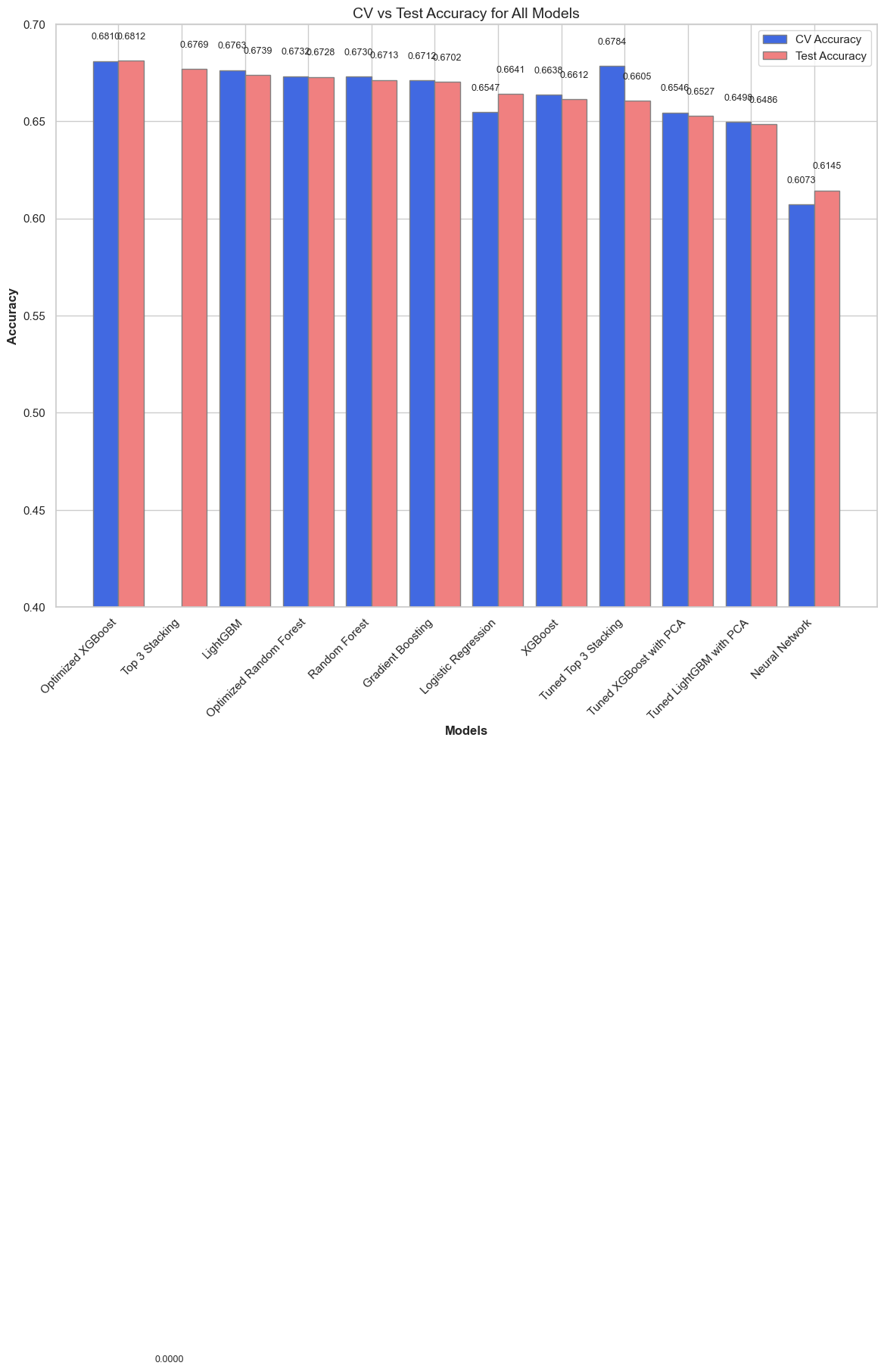
****

Figure 8.1. Cross-Validation and Test Accuracy Results

Here we plot the confusion matrices of the top two models. The confusion matrices of the rest models can be found in Appendix A.

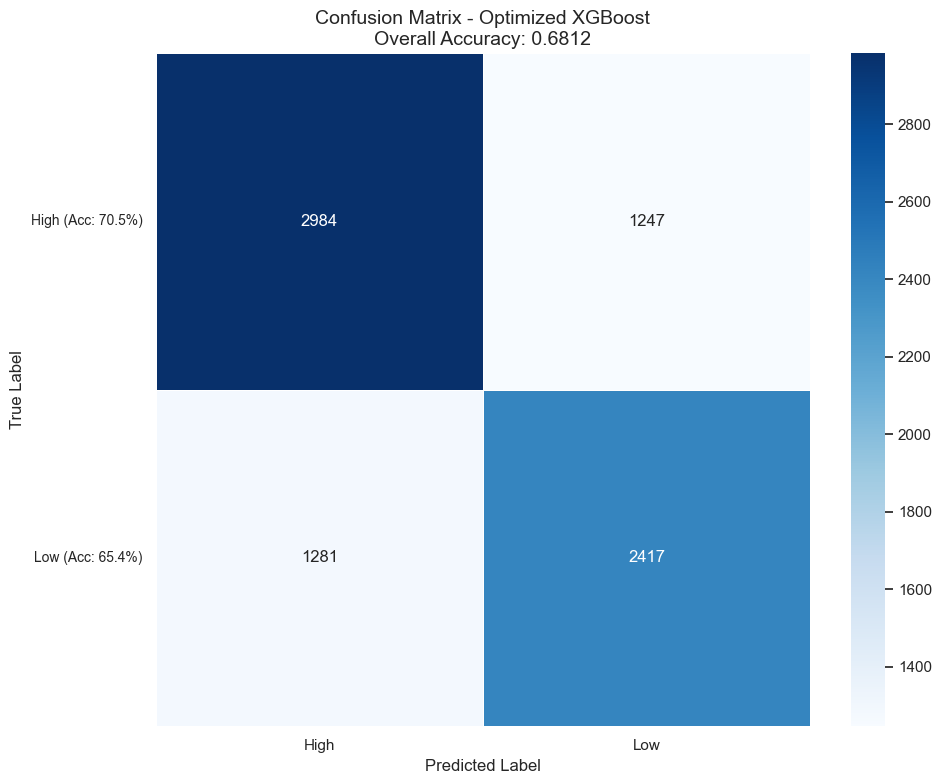


Figure 8.2. Confusion matrix of the Optimized XGBoost Model

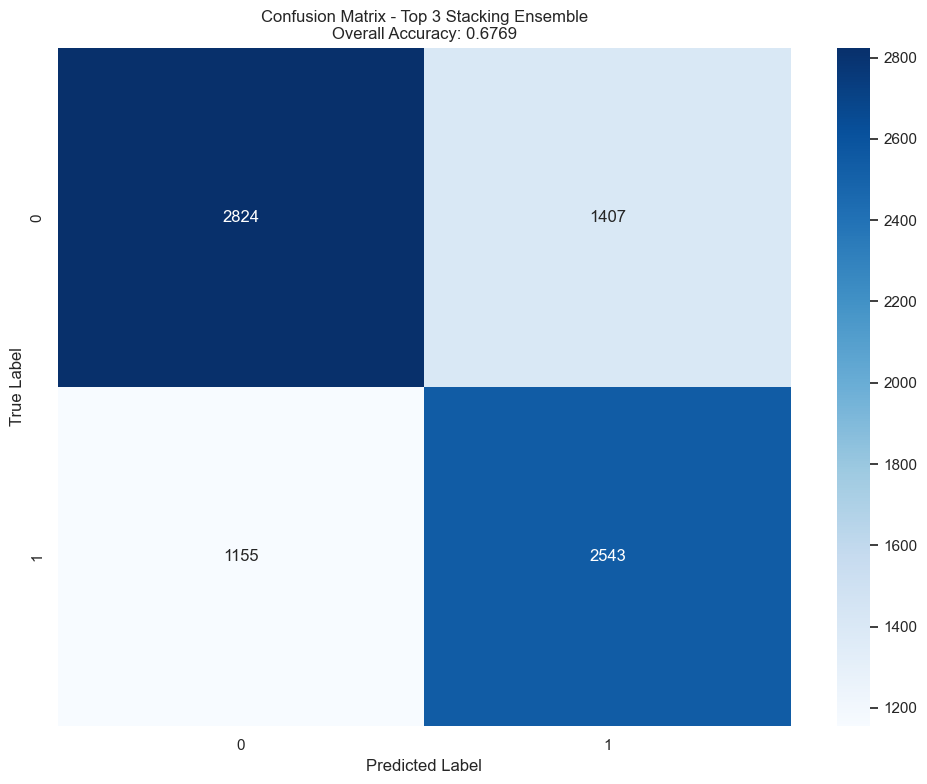


Figure 8.3. Confusion matrix of the Stacking model

We find that of the two models listed, the Optimized XGBoost Model predicts better for the High class, while the Stacking Model predicts better for the Low class. This actually gives us a clearer direction for future research, which means that the two models can be utilized again to continue training new Stacking models.

The test accuracy for most models closely aligns with CV accuracy, indicating that the models did not suffer from significant overfitting.

We also calculated the feature importance of the data from our best model. The XGBoost library has the function to directly output feature importance.

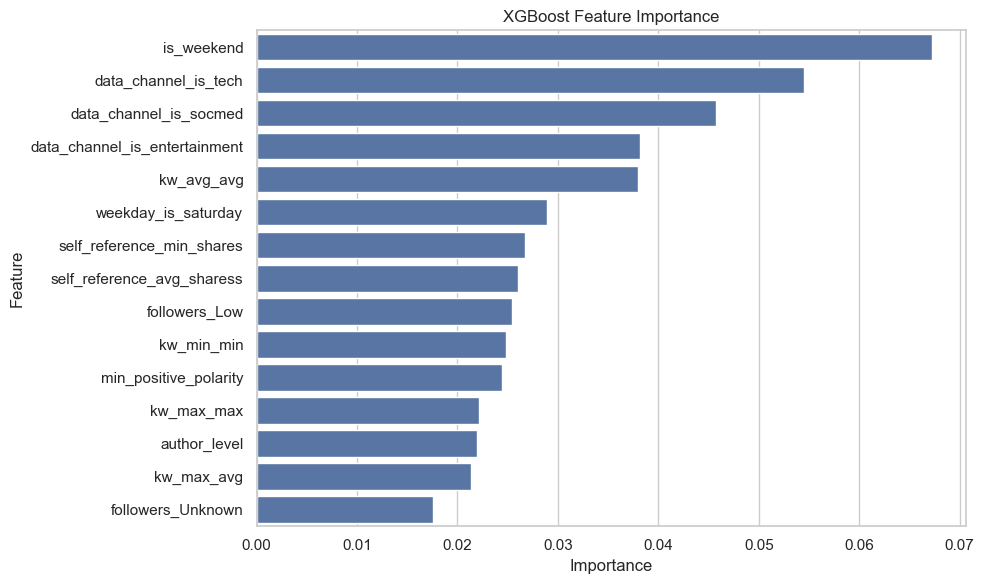


Figure 8.4. Feature Importance of XGBoost

The most significant feature we found was whether or not the article was published over the weekend. This is very intuitive. We suggest that if the frequency of posting articles is once or twice a week, then authors should try to choose to post on weekends.

As mentioned previously, we attempted to extract components that contribute 95% of the variance using the PCA method. Even though our final best trained model did not use PCA but the full features, we still plotted the change in total variance contribution with the PCA process.

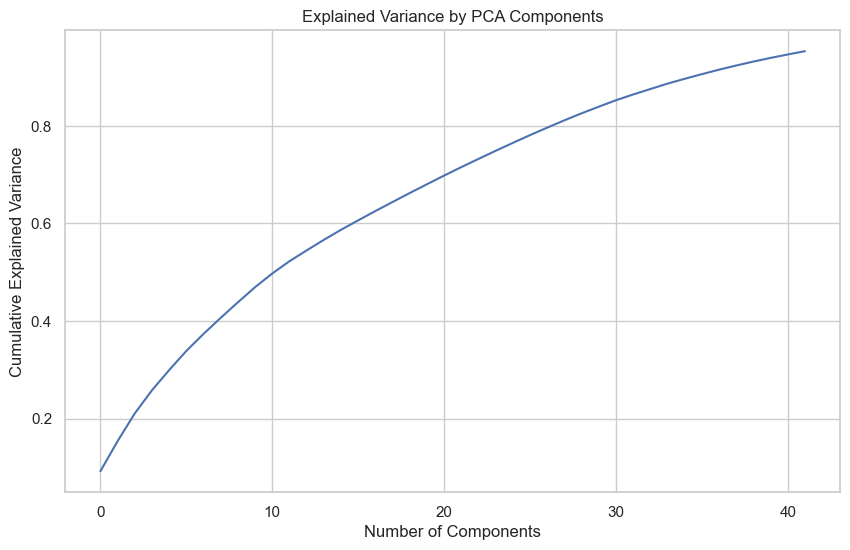


Figure 8.5. Explained Variance by PCA Components

We found that the first 42 principal components with the highest variance contribution contribute 95% of the variance. There are only 65 features in total, indicating that the original data is indeed very complex and the multicollinearity between features is weak.

From a business perspective, accurately predicting shareability can help media organizations optimize content creation efforts, improve engagement rates, and increase advertising revenue. By prioritizing articles that are likely to gain traction, publishers can allocate resources more efficiently, reducing the time spent on articles with lower engagement potential. For example, if an article is predicted to be highly shareable, editors can promote it more aggressively, resulting in increased traffic and higher ad revenue.

Beyond media, this predictive framework has broader applications. The methodology could extend to social media marketing, product recommendations, and online ad targeting, where predicting content virality is crucial for audience engagement. Future work may involve integrating real-time engagement metrics, such as early reader interactions, into the prediction model, further refining its accuracy and business utility.

### **Chapter 9: Lessons Learned and Recommendations**

This project we have done demonstrated that machine learning models can effectively predict online news popularity, but there are several areas for improvement. One key lesson is that structured datasets like the one we used often favor tree-based models such as XGBoost and LightGBM, while deep learning models may require more complex architectures or additional data to improve performance. Additionally, feature selection and transformation techniques significantly impacted model accuracy, highlighting the importance of preprocessing steps in predictive modeling.

For future improvements, integrating advanced deep learning models such as Transformers or Graph Neural Networks (GNNs) may enhance predictive performance by capturing textual and relational features more effectively. These models could better understand the linguistic patterns in article text and user interactions, leading to stronger predictions.

Beyond model enhancement, incorporating third-party datasets could add valuable context to the analysis. Social media engagement metrics from platforms like Twitter, Facebook, and Reddit could provide real-time indicators of virality. Additionally, Google Trends data could be used to measure public interest in article topics at the time of publication, offering another predictive variable. These external datasets would allow the model to learn from broader audience behavior rather than relying solely on article attributes.

Another recommendation is to deploy the predictive model in a real-time content management system (CMS), allowing media organizations to dynamically adjust their content strategies. A web-based dashboard could provide real-time insights on article shareability, enabling editors and marketers to optimize headlines, images, and keywords for maximum reach. Furthermore, the system could suggest modifications to articles before publishing, increasing the likelihood of high engagement.

Overall, while our study has some insights into news article popularity prediction, future work should focus on integrating richer datasets, experimenting with advanced deep learning models, and developing automated decision-support systems. These enhancements will ensure that machine learning continues to play a pivotal role in media strategy optimization.

### **Appendix A: Results of Models**

| **Rank** | **Model** | **CV Accuracy** | **Test Accuracy** | **Test F1 Score** |
| --- | --- | --- | --- | --- |
| 1 | Optimized XGBoost | 0.6810 | 0.6812 | 0.6811 |
| 2 | Top 3 Stacking | 0.0000 | 0.6769 | 0.6772 |
| 3 | LightGBM | 0.6763 | 0.6739 | 0.6735 |
| 4 | Optimized Random Forest | 0.6732 | 0.6728 | 0.6729 |
| 5 | Random Forest | 0.6730 | 0.6713 | 0.6709 |
| 6 | Gradient Boosting | 0.6712 | 0.6702 | 0.6703 |
| 7 | Logistic Regression | 0.6547 | 0.6641 | 0.6634 |
| 8 | XGBoost | 0.6638 | 0.6612 | 0.6611 |
| 9 | Tuned Top 3 Stacking | 0.6784 | 0.6605 | 0.6599 |
| 10 | Tuned XGBoost with PCA | 0.6546 | 0.6527 | 0.6528 |
| 11 | Tuned LightGBM with PCA | 0.6498 | 0.6486 | 0.6490 |
| 12 | Neural Network | 0.6073 | 0.6145 | 0.6146 |

Table A.1. Metrics of All Models

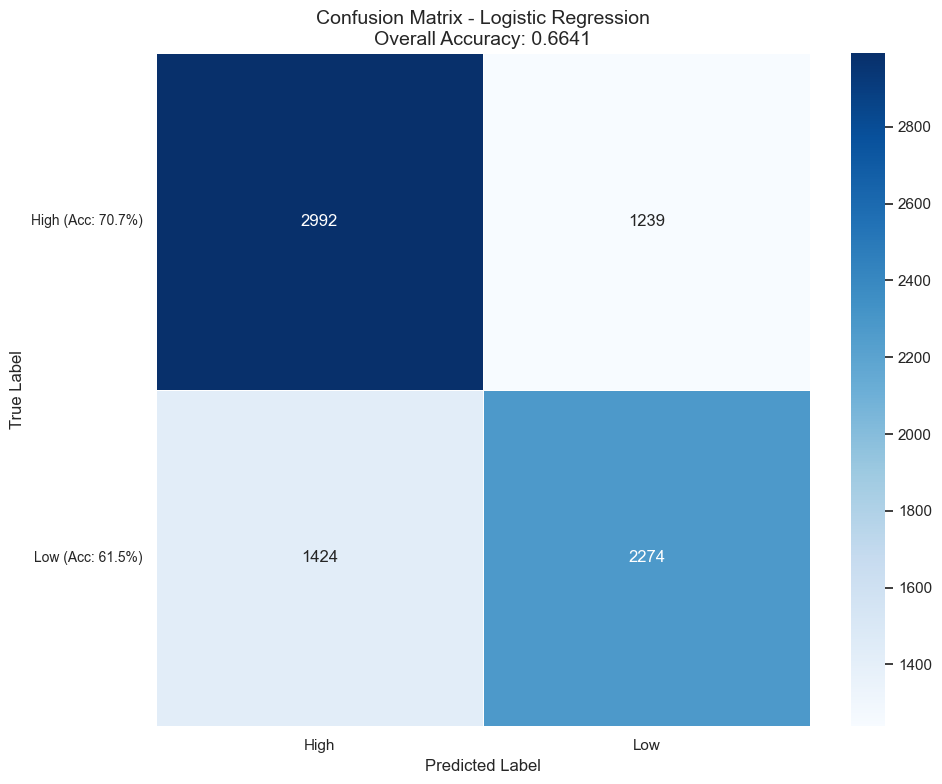


Figure A.1. Confusion Matrix of Logistic Regression Model

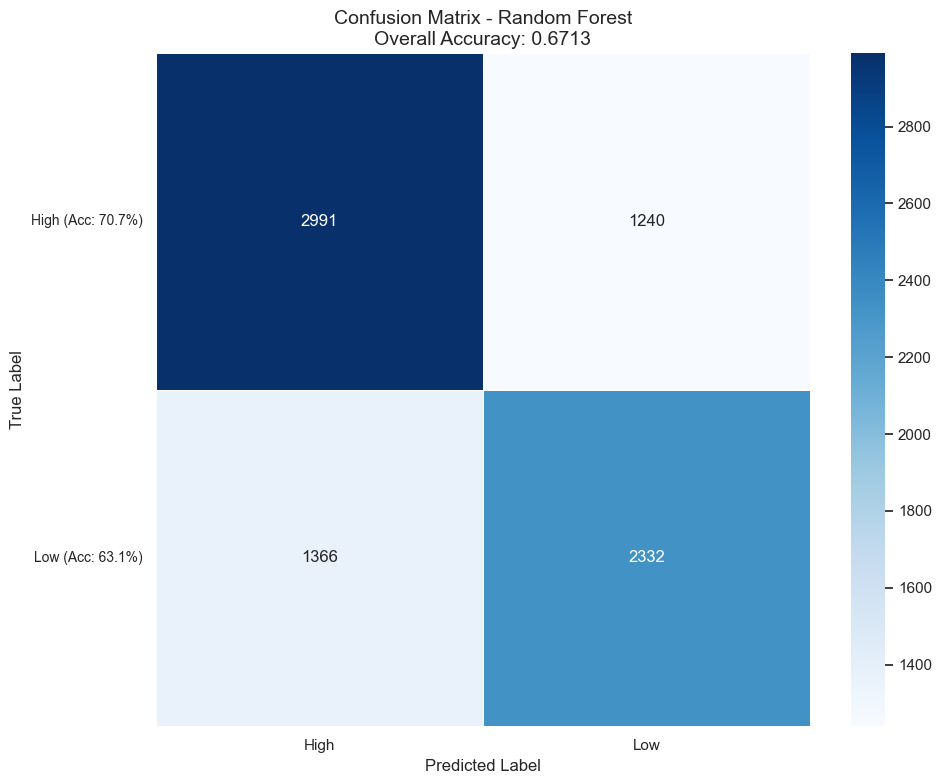


Figure A.2. Confusion Matrix of Random Forest Model

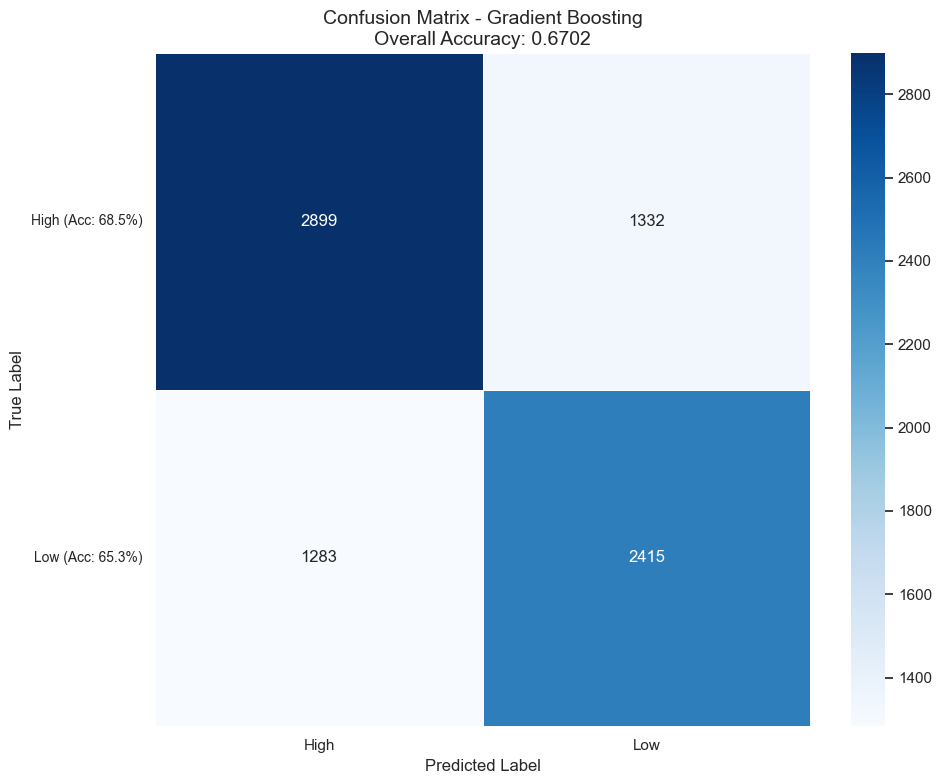


Figure A.3. Confusion Matrix of Gradient Boosting Model

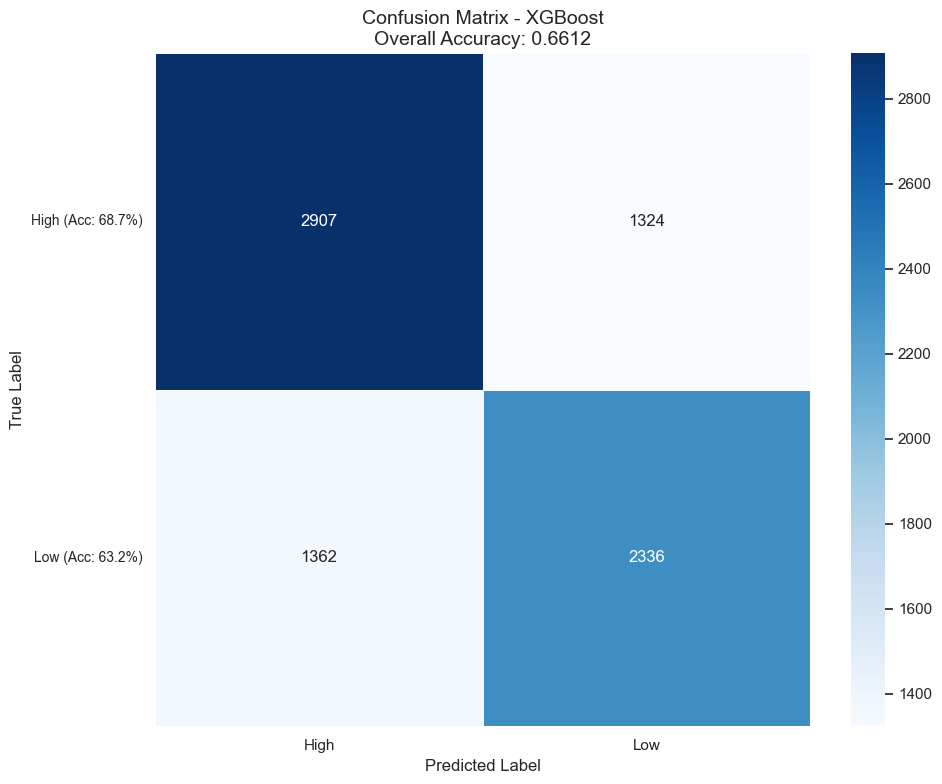


Figure A.4. Confusion Matrix of XGBoost Model without Optimization

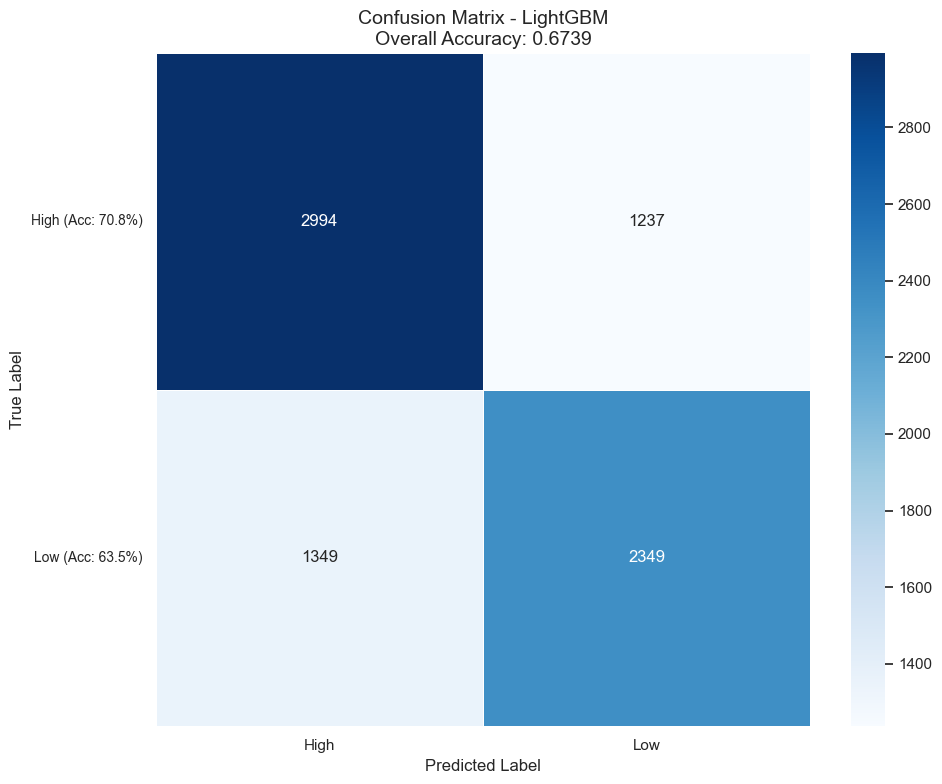


Figure A.4. Confusion Matrix of LightGBM Model

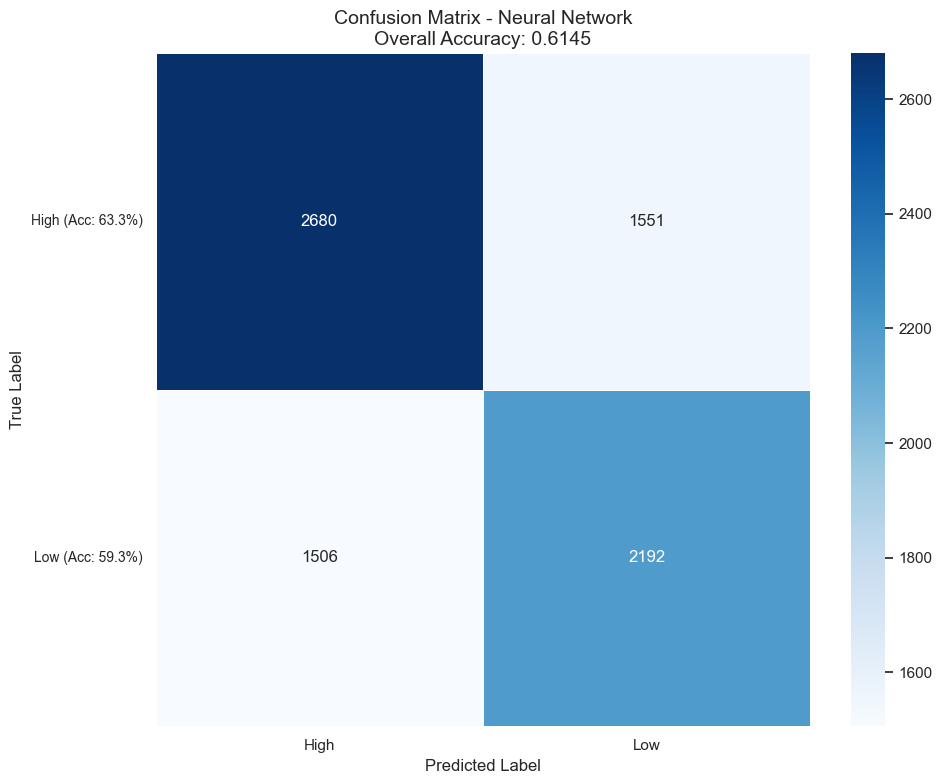


Figure A.5. Confusion Matrix of MLP

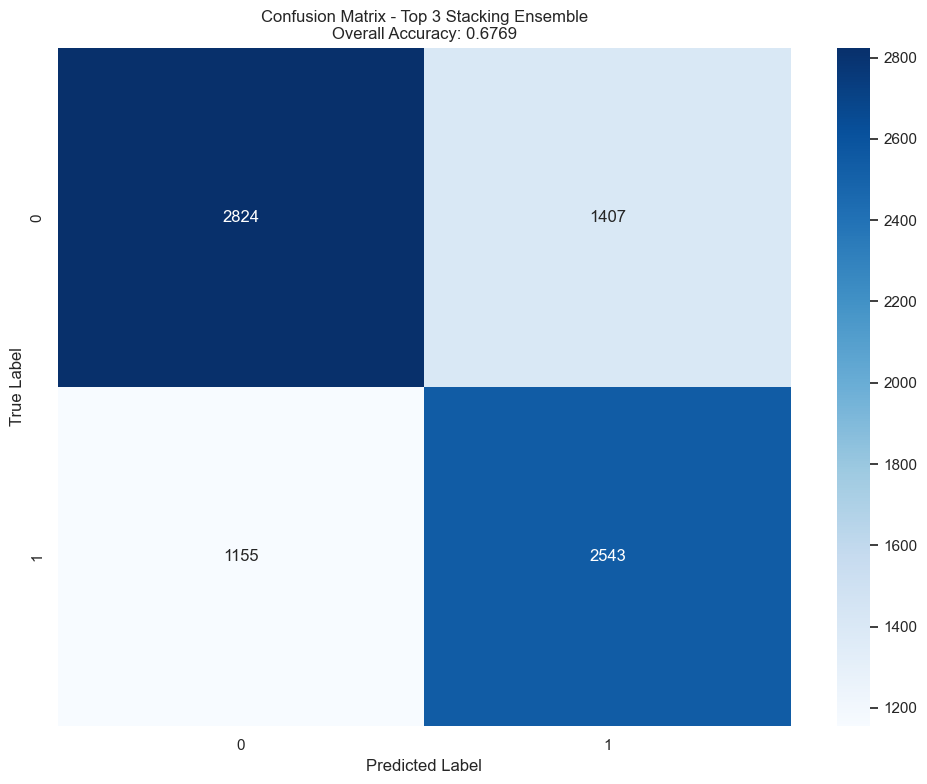


Figure A.6. Confusion Matrix of Stacking Model

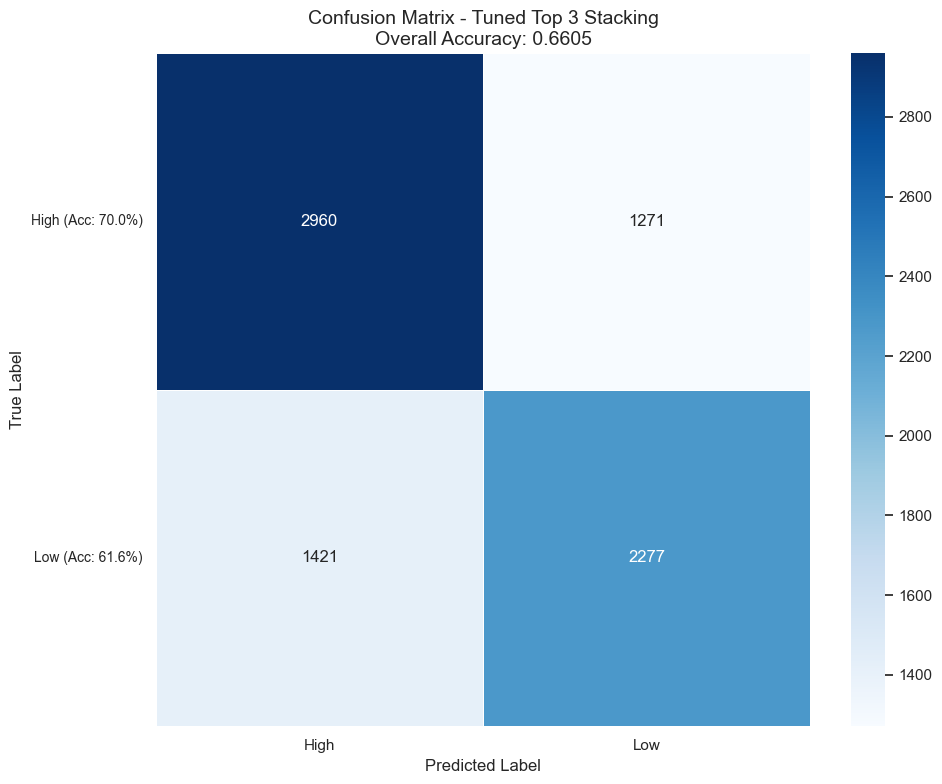


Figure A.7. Confusion Matrix of Tuned Stacking Model

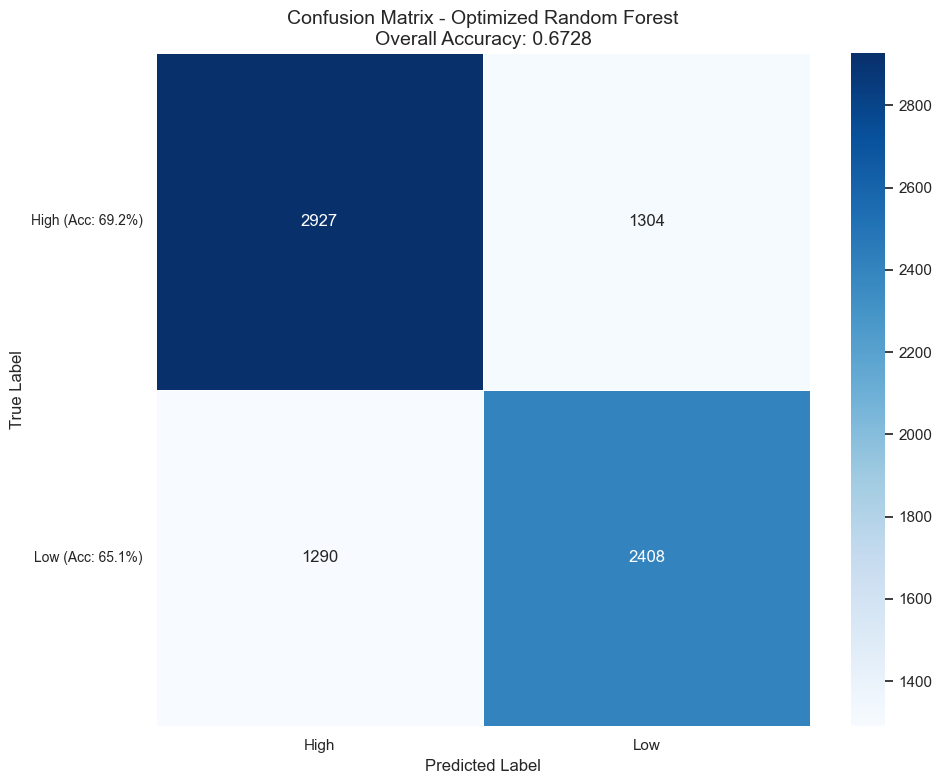


Figure A.8. Confusion Matrix of Optimized Random Forest Model

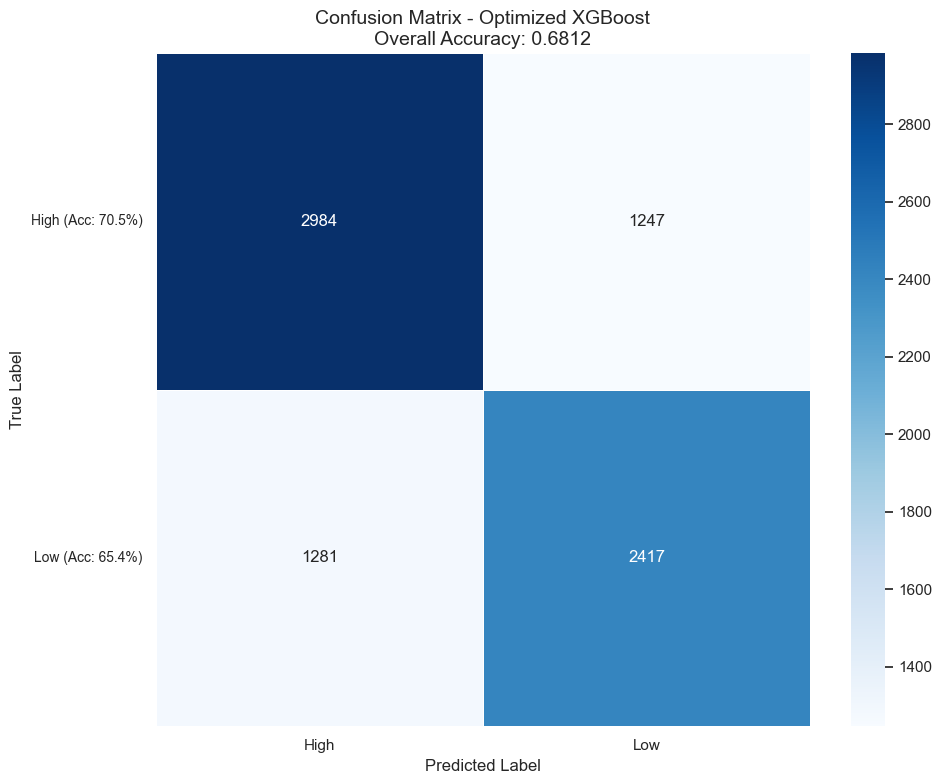


Figure A.9. Confusion Matrix of Optimized XGBoost Model

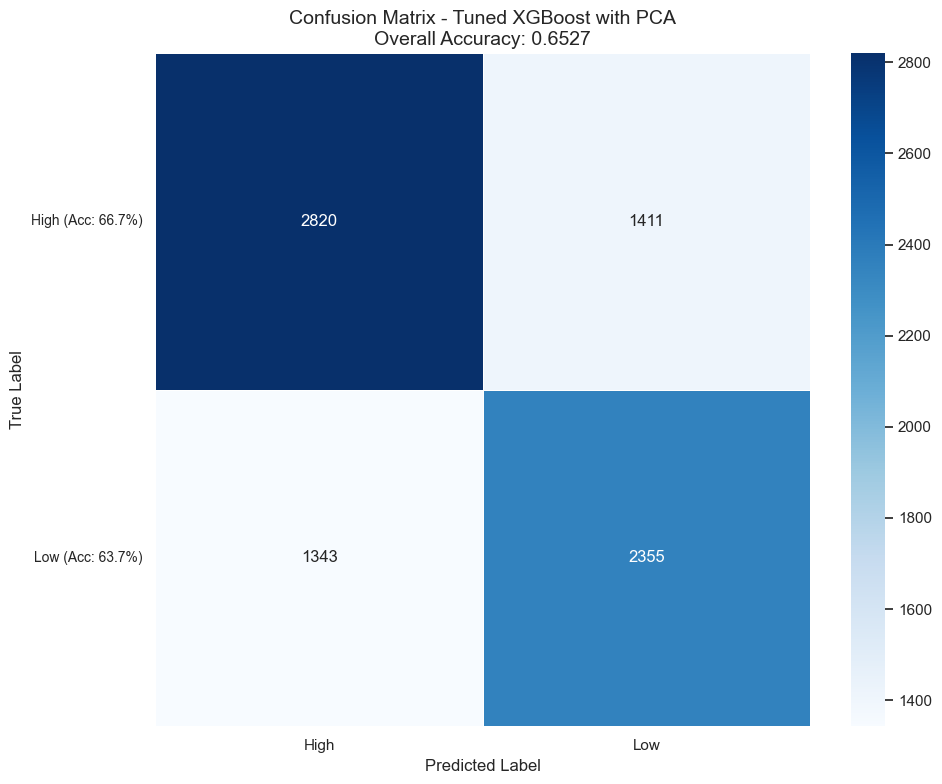


Figure A.10. Confusion Matrix of XGBoost Model with PCA

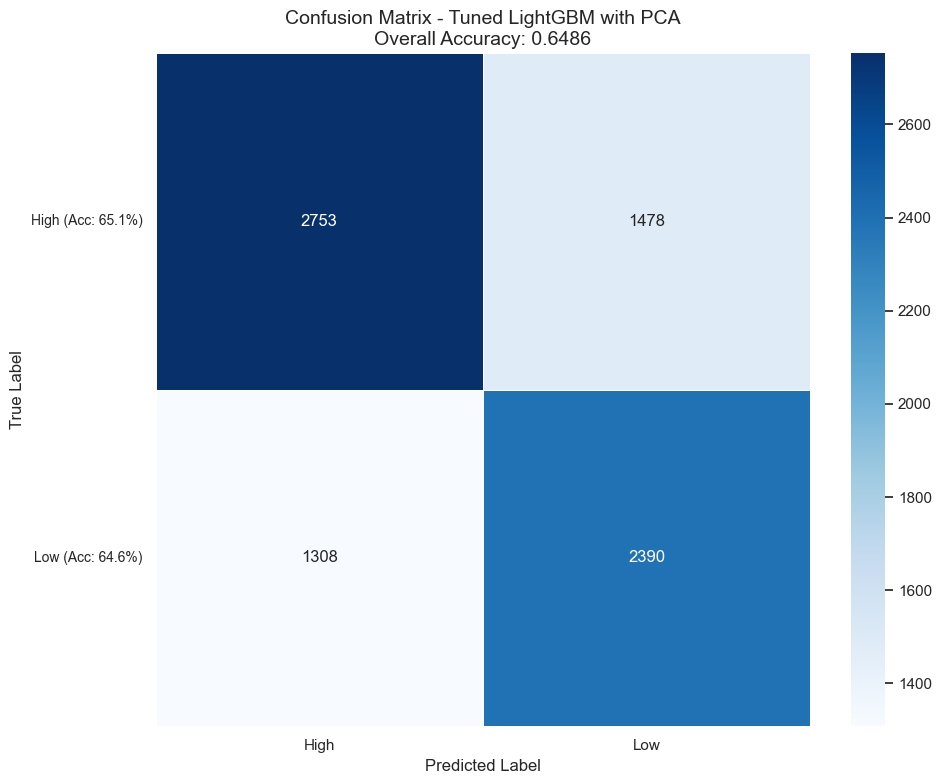


Figure A.10. Confusion Matrix of XGBoost Model with PCA

### **References**

Arora, Anuja, et al. "*A Novel Multimodal Online News Popularity Prediction Model Based on Ensemble Learning*." Expert Systems, vol. 40, no. 8, 2023, <https://doi.org/10.1111/exsy.13336> .

Fernandes, Kelwin, et al. "*A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*." Progress in Artificial Intelligence, vol. 9273, 2015, pp. 535-546, <https://doi.org/10.1007/978-3-319-23485-4_53> .

Gao, Ge, et al. "A Novel Multimodal Online News Popularity Prediction Model Based on Deep Residual Neural Network." Expert Systems 39, no. 4 (2022): e13336. <https://doi.org/10.1111/exsy.13336>

Gebhard, Lukas, and Felix Hamborg. "The POLUSA Dataset: 0.9M Political News Articles Balanced by Time and Outlet Popularity." arXiv preprint arXiv:2005.14024 (2020). <https://arxiv.org/abs/2005.14024>

Qi, Tao, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. "PP-Rec: News Recommendation with Personalized User Interest and Time-Aware News Popularity." arXiv preprint arXiv:2106.01300 (2021). <https://arxiv.org/abs/2106.01300>

Rajagopal, Suharshala, et al. "*Online News Popularity Prediction before Publication: Effect of Readability, Emotion, Psycholinguistics Features*." IAES International Journal of Artificial Intelligence, vol. 11, no. 2, 2022, pp. 539-545, <https://doi.org/10.11591/ijai.v11.i2.pp539-545> .

Roy, Sayar Ghosh, et al. "Towards Proactively Forecasting Sentence-Specific Information Popularity within Online News Documents." arXiv preprint arXiv:2301.00152 (2023). <https://arxiv.org/abs/2301.00152>

Shang, Yingdan, et al. "*Predicting the Popularity of Online Content by Modeling the Social Influence and Homophily Features*." Frontiers in Physics, vol. 10, 2022, 915756, <https://doi.org/10.3389/fphy.2022.915756> .

Tsai, Min-Jen, and You-Qing Wu. "*Predicting Online News Popularity Based on Machine Learning*." Computers and Electrical Engineering, vol. 102, 2022, 108198, <https://doi.org/10.1016/j.compeleceng.2022.108198> .