Predicting News Deserts Using Supervised Machine Learning

ABSTRACT

The decline of local newspapers has led to the emergence of news deserts – areas lacking access to critical local information – posing a threat to community engagement and democracy. This study aims to predict which U.S. counties are most at risk of becoming news deserts by developing machine learning models based on socioeconomic, geographic, and circulation data. Addressing class imbalance and data noise, we employed classifiers such as Logistic Regression, Random Forest, XGBoost, Support Vector Machines, K-Nearest Neighbors, and Naive Bayes, combined with resampling techniques like SMOTE, Tomek Links, SMOTETomek, SMOTEENN, and ADASYN. Our analysis found that XGBoost combined with ADASYN performed best, achieving an F2-Score of 0.486 and AUC-PR of 0.467 on test data. These results provide valuable insights for policymakers aiming to develop targeted interventions to preserve local media ecosystems and strengthen democratic processes.

INTRODUCTION

The emergence of news deserts – communities with limited or no access to reliable, comprehensive local news – has raised significant concerns about the health of democracy and community engagement in the United States (Franklin, 2014; Abernathy, 2018). The decline of local newspapers, often attributed to economic pressures, shifts in advertising revenue, and the rise of digital media platforms, has left a substantial portion of the population without essential information sources (Ali, 2017; Abernathy & Franklin, 2022). Estimates suggest that nearly one-fifth of the country's population resides in news deserts (Abernathy, 2020), highlighting the urgency of addressing this issue.

While existing research has mapped current news deserts and identified correlating factors (Napoli et al., 2018; Hindman, 2018; Stonebely, 2023), there is a gap in predictive modeling to anticipate future at-risk areas. This study aims to fill that gap by developing a predictive model to identify U.S. counties at high risk of becoming news deserts. Leveraging machine learning algorithms and incorporating socioeconomic variables along with spatial neighbor data, we seek to provide a proactive tool for stakeholders to identify, intervene, and support local journalism in vulnerable communities.

METHODOLOGY & DATA

An important consideration in our methodology is the imbalanced nature of the overall dataset from where we create the training model - meaning our data has a far higher number of *non-news desert* counties, and thus more of their attributes, as compared to *news deserts*. Classification algorithms, however, focus on the well-represented, or majority class, traditionally (Sawangarreerak & Thanathamathee, 2020). To rectify the class imbalance problem, we use different resampling methods: ADASYN, oversampling using Synthetic Minority Over-sampling Technique (SMOTE), undersampling using Tomek links, and combine both (SMOTE + Tomek) methods. The different classification models employed are retrained on the resampled training data, which then enables them to find patterns and relationships in the test data with an adjusted focus due to the newly synthesized minority class instances. This helps in improving the models' ability to classify minority class instances correctly.

Our news data is obtained from the 'State of Local News' report published in 2022 (Abernathy & Franklin, 2022). The second part of our method incorporates county-level metadata, and we use Federal Information Processing Standard code (FIPS) to anchor counties without any local news outlets in our dataset. Once we specifically map all *news deserts* using FIPS codes, it enables us to expand our analysis to utilize several different open datasets with different information. Broadly, the data we look at are population, race, age, educational attainment, median household income, broadband access, voting, and county GDP. We obtain the information from different datasets, including the ACS 5-year Census data, MIT Election Data, USDA Economic Research Service, and the U.S. Bureau of Economic Analysis.

Missing values can bias machine learning models and degrade predictive performance (Little & Rubin, 2019). In this study, missing values in the socioeconomic variables were imputed using the median imputation strategy. The median is robust to outliers and provides a central tendency measure that is less affected by skewed data distributions (Gelman et al., 2013). Formally, for a variable X with missing entries, each missing value X_i is replaced by the median X-bar of the observed values:

X_i = X-bar if X_i is missing

A news desert label is based on there being no local newspapers in circulation at the county level. The target variable Y_i, news desert county, is defined as:

$$Y_i = \begin{cases} 1 \; if \; county \; is \; a \; news \; desert \\ 0 \; otherwise \end{cases}$$

To capture the complex interplay between socioeconomic factors, we also use two interaction terms:

Population Density x GDP,

Income x Broadband.

These interaction terms help model non-linear relationships and interactions between variables, which are essential in capturing the multifaceted nature of news desert formation.

For a given county *i*, the population-weighted average of a socio-economic feature x from its neighboring counties N_i is:

$$x_{\text{neighbor\_avg}}^{(i)} = \frac{\sum_{j \in N(i)} x^{(j)} \cdot P(j)}{\sum_{j \in N(i)} P(j)}$$

where:

- $x^{(j)}$ is the value of a given feature x in neighboring county j,
- $P^{(j)}$ is the total population of county j,
- N_i is the set of neighboring counties of county i.

This allows us to incorporate the influence of neighboring counties' socio-economic conditions on the local news environment of any given county i.

RESULTS

An initial evaluation using 5-fold stratified cross-validation revealed that the XGBoost classifier combined with the SMOTEENN resampling technique achieved the highest mean F2-score of 0.491. The Random Forest classifier (with SMOTEENN) and XGBoost (with SMOTE) also demonstrated solid performance, with a mean F2-score of 0.439 and 0.435 respectively. Table 1 displays the five highest F2-scores obtained after cross-validation.

While the XGBoost model with SMOTEENN achieved the highest cross-validation F2-score, cross-validation performance does not always guarantee the best performance on unseen data due to potential overfitting or variance in data distribution (Browne, 2000). Therefore, we evaluate the models on the test set to determine their generalization capabilities overall. Once this step concluded, we found that the XGBoost classifier with ADASYN resampling achieved the best balance of metrics: the highest F2-score of 0.485, a recall of 0.488, and a precision of 0.476, demonstrating its effectiveness in identifying at-risk counties while maintaining a balance between precision and recall (see Tables 2 - 7). In contrast, models like Logistic Regression performed less effectively despite resampling, due to their inability to capture non-linear relationships without extensive feature engineering. The superior performance of ensemble methods like XGBoost and Random Forest indicates the importance of capturing complex patterns in the data.

Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves are plotted to visualize the models' discriminative abilities. The XGBoost model with ADASYN demonstrated favorable performance, with an ROC-AUC of 0.86 and an AUC-PR of 0.47 (see

Figure 1 and 2). These curves reinforce the model's capability to distinguish between at-risk and non-at-risk counties effectively. We also generate confusion matrices for each combination of classifier and resampling method (a few examples Table 8 - 15). The matrices provide more substantive insights into the number of true positives, false positives, true negatives, and false negatives for each model, allowing us to assess how well each classifier distinguishes between counties at-risk that are news deserts and non-news deserts.

Focusing on our initial hypothesis RQ1 and RQ2, our study demonstrates that advanced machine learning techniques can effectively predict at-risk counties even in the presence of significant class imbalance. The utilization of ensemble methods capable of capturing complex, nonlinear relationships among different socioeconomic variables enhances predictive accuracy and utilization of various resampling techniques further improved model performance, allowing the model to learn the underlying patterns more effectively.

Resampling techniques, particularly ADASYN, also significantly improved the model's ability to identify at-risk counties by balancing the class distribution and focusing on difficult-to-learn instances. The XGBoost classifier with ADASYN achieved the highest F2-score on the test data, indicating enhanced identification of news deserts while maintaining a balanced performance in terms of precision (0.476) and recall (0.488), which displays its effectiveness in handling imbalanced datasets. The model's approach demonstrates the practical utility of advanced machine learning models in handling geographically diverse datasets with imbalanced classes.

For RQ3, we explored the effect of interaction terms and including socioeconomic characteristics from neighboring counties in addition to within-county features. We had also calculated population-weighted averages of socioeconomic variables from adjacent counties and incorporated these into our feature set. The feature importance analysis revealed that features like population density – GDP interaction, income – broadband interaction, average broadband access of neighboring counties, and average GOP vote percentage of neighboring counties were also important predictors. By integrating these influences, the model captured spatial patterns and dependencies, leading to improved predictive performance in modeling news desertification.

An examination of the feature importances from the XGBoost model revealed that the interaction term 'Population Density × GDP' was the most significant predictor, accounting for approximately 19% of the model's decision-making process (see Table 16). This finding underscores the hypothesis that the combined effect of economic activity and population concentration critically influences the sustainability of local news outlets. Counties with low population density and GDP may lack the necessary advertising revenue and subscriber base to support local newspapers. The other top five features included:

- Hispanic/Latino Population Percentage: Higher percentages correlate with increased risk, potentially due to historical underrepresentation and economic disparities in these communities;
- GDP: Lower GDP values were associated with higher risk, emphasizing the role of economic vitality in sustaining local news outlets;
- Broadband access: Limited broadband access emerged as a significant factor, underscoring the importance of digital infrastructure in modern news dissemination;
- Population Density: Lower population densities were linked to higher at- risk counties, reflecting challenges in sustaining newspapers in sparsely populated areas;

Using our best-performing XGBoost model with ADASYN, we also predicted probabilities for counties, currently not classified as news deserts, which are predicted to be at-risk (Table 17). These high-risk counties typically share characteristics such as low population density, economic distress, limited broadband access, and higher percentages of minority populations.

By addressing these research questions, this study contributes to a deeper understanding of the factors influencing news desertification and demonstrates the practical utility of machine learning models in informing policy interventions. The identification of these counties along with the relevant actionable insights can provide a platform for policymakers and stakeholders to implement targeted interventions to support local journalism in vulnerable areas, and, by extension, fortify the democratic process.

Tables & Figures

Table 1: Cross – Validation Results

Classifier	Resampling Method	F2-Score
XGBoost	SMOTEENN	0.491
Random Forest	SMOTEENN	0.439
XGBoost	Smote	0.436
Random Forest	None	0.43
XGBoost	ADASYN	0.428

Table 2

Model	**Resampling**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**F2-Score**	**ROC-AUC**	**Average Precision (AUC-PR)**
Logistic Regression	None	0.933	0.667	0.049	0.091	0.060	0.751	0.273
Logistic Regression	Smote	0.666	0.115	0.585	0.193	0.323	0.736	0.279
Logistic Regression	ADASYN	0.644	0.116	0.634	0.195	0.334	0.734	0.245
Logistic Regression	SMOTETomek	0.666	0.115	0.585	0.193	0.323	0.737	0.282
Logistic Regression	SMOTEENN	0.522	0.106	0.805	0.187	0.347	0.730	0.231
Logistic Regression	Tomek Links	0.933	0.667	0.049	0.091	0.060	0.749	0.272

Table 3

Model	Resampling	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Average Precision (AUC-PR)
Random Forest	None	0.943	0.818	0.220	0.346	0.257	0.870	0.545
Random Forest	Smote	0.915	0.375	0.366	0.370	0.368	0.844	0.435
Random Forest	ADASYN	0.927	0.462	0.439	0.450	0.443	0.858	0.471
Random Forest	SMOTETomek	0.928	0.475	0.463	0.469	0.466	0.832	0.453
Random Forest	SMOTEENN	0.880	0.282	0.488	0.357	0.426	0.805	0.391
Random Forest	Tomek Links	0.942	0.714	0.244	0.364	0.281	0.864	0.530

Table 4

Model	Resampling	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Average Precision (AUC-PR)
XGBboost	None	0.947	0.765	0.317	0.448	0.359	0.899	0.581
XGBboost	Smote	0.922	0.432	0.463	0.447	0.457	0.872	0.502
XGBboost	ADASYN	0.928	0.476	0.488	0.482	0.485	0.858	0.466
XGBboost	SMOTETomek	0.915	0.386	0.415	0.400	0.409	0.840	0.441
XGBboost	SMOTEENN	0.889	0.309	0.512	0.385	0.453	0.829	0.403
XGBboost	Tomek Links	0.953	0.842	0.390	0.533	0.437	0.888	0.632

Table 5

Model	Resampling	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Average Precision (AUC-PR)
SVM (RBF Kernel)	None	0.834	0.208	0.512	0.296	0.396	0.764	0.224
SVM (RBF Kernel)	Smote	0.779	0.152	0.488	0.231	0.338	0.711	0.232
SVM (RBF Kernel)	ADASYN	0.767	0.154	0.537	0.239	0.358	0.710	0.218
SVM (RBF Kernel)	SMOTETomek	0.779	0.152	0.488	0.231	0.338	0.712	0.233
SVM (RBF Kernel)	SMOTEENN	0.729	0.145	0.610	0.235	0.372	0.710	0.190
SVM (RBF Kernel)	Tomek Links	0.832	0.206	0.512	0.294	0.395	0.765	0.225

Table 6

Model	Resampling	Accuracy	Precision	Recall	F1-Score	F2-Score	ROC-AUC	Average Precision (AUC-PR)
KNN	None	0.933	0.600	0.073	0.130	0.089	0.676	0.235
KNN	Smote	0.799	0.155	0.439	0.229	0.321	0.665	0.150
KNN	ADASYN	0.784	0.138	0.415	0.207	0.296	0.675	0.160
KNN	SMOTETomek	0.799	0.155	0.439	0.229	0.321	0.667	0.150
KNN	SMOTEENN	0.724	0.121	0.488	0.194	0.304	0.662	0.113
KNN	Tomek Links	0.933	0.600	0.073	0.130	0.089	0.674	0.230

Table 7

Model	**Resampling**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**F2-Score**	**ROC-AUC**	**Average Precision (AUC-PR)**
Naive Bayes	None	0.271	0.079	0.902	0.145	0.291	0.676	0.154
Naive Bayes	Smote	0.290	0.080	0.902	0.148	0.296	0.665	0.148
Naive Bayes	ADASYN	0.283	0.080	0.902	0.147	0.295	0.656	0.143
Naive Bayes	SMOTETomek	0.288	0.080	0.902	0.147	0.296	0.665	0.148
Naive Bayes	SMOTEENN	0.286	0.080	0.902	0.147	0.296	0.667	0.142
Naive Bayes	Tomek Links	0.273	0.081	0.927	0.148	0.299	0.676	0.154

Table 8

Confusion Matrix for Logistic Regression with SMOTEENN Resampling

Actual	Predicted	
	Negative	Positive
	(Non-news desert)	(News Desert)
Negative (Non-news desert)	281	279
Positive (News Desert)	8	33

Table 9

Confusion Matrix for Random Forest with SMOTETomek Resampling

Actual	Predicted	
	Negative	Positive
	(Non-news desert)	(News Desert)
Negative (Non-news desert)	539	21
Positive (News Desert)	22	19

Table 10

Confusion Matrix for Random Forest with SMOTETomek Resampling

Actual	Predicted	
	Negative	Positive
	(Non-news desert)	(News Desert)
Negative (Non-news desert)	539	21
Positive (News Desert)	22	19

Table 11

Confusion Matrix for XGBoost with ADASYN Resampling

Actual	Predicted	
	Negative	Positive
	(Non-news desert)	(News Desert)
Negative (Non-news desert)	538	22
Positive (News Desert)	21	20

Table 13

Confusion Matrix for SVM (RBF Kernel) with Tomek Links Resampling

Actual	Predicted	
	Negative	Positive
	(Non-news desert)	(News Desert)
Negative (Non-news desert)	479	81
Positive (News Desert)	20	21

Table 14

Confusion Matrix for KNN with SMOTETomek Resampling

Actual	Predicted	
	Negative (Non-news desert)	Positive (News Desert)
Negative (Non-news desert)	462	98
Positive (News Desert)	23	18

Table 15

Confusion Matrix for Naive Bayes without Resampling

Actual	Predicted	
	Negative (Non-news desert)	Positive (News Desert)
Negative (Non-news desert)	126	434
Positive (News Desert)	4	37

Table 16

Feature	Importance
Population Density * GDP	0.194492
Hispanic/Latino	0.101181
GDP (USD)	0.073908
Broadband Access	0.045534
Population Density (Per Sq. Mile)	0.044036
Black (Neighboring County Average)	0.042858
Median HH Income * Broadband	0.042379
GDP (Neighboring County Average)	0.040427
Political Affiliation (GOP)	0.040286
Age 65+ (Neighboring County Average)	0.03915
Broadband Access (Neighboring County Average)	0.038789
Age 65+	0.038114
Black	0.03733
Political Affiliation GOP (Neighboring County Average)	0.035725
Median HH Income (Neighboring County Average)	0.035059
Median Household Income	0.033871
Bachelor's Degree (Neighboring County Average)	0.033841
Bachelor's Degree (Pct)	0.031365
Hispanic/Latino (Neighboring County Average)	0.028391
Population Density (Neighboring County Average)	0.023266

Table 17

FIPS	County	State	County At-Risk Probability
48137	Edwards	TX	0.995
51079	Greene	VA	0.971
46007	Bennett	SD	0.97
24029	Kent	MD	0.964
16077	Power	ID	0.948
19087	Henry	IA	0.947
13167	Johnson	GA	0.923
37177	Tyrrell	NC	0.904
29135	Moniteau	MO	0.903
35047	Sandoval	NM	0.894

XGBoost – ADASYN model prediction of top ten counties at-risk

Figure 1

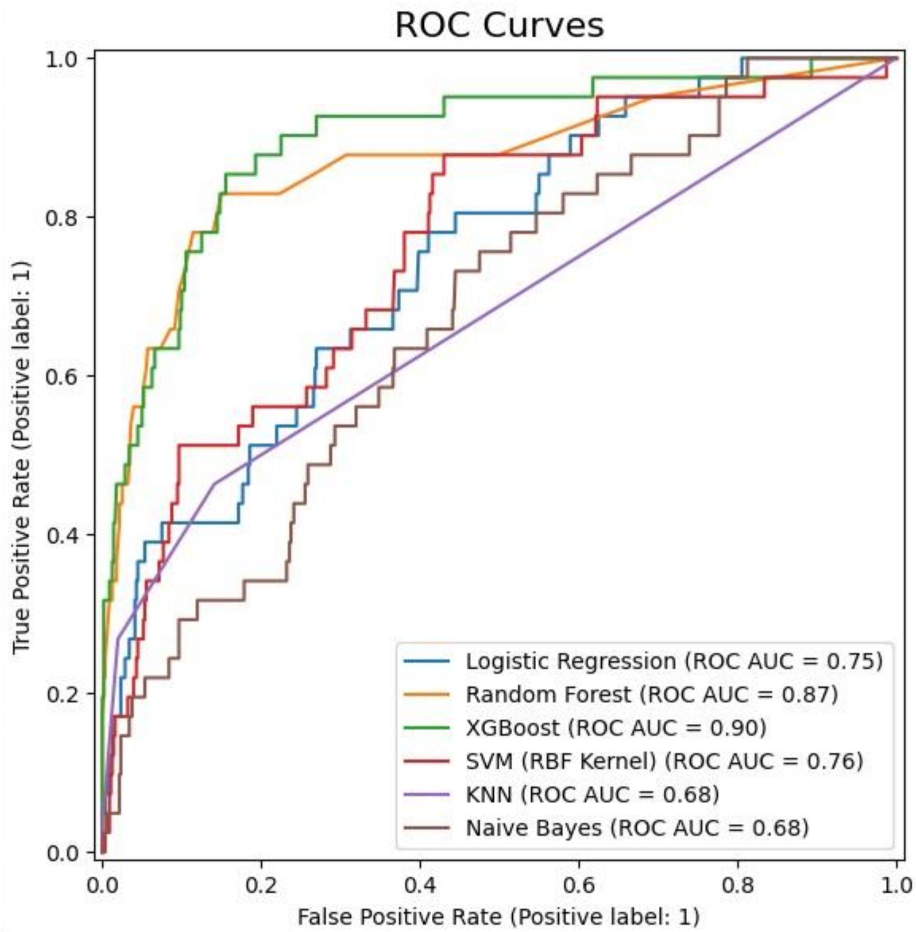
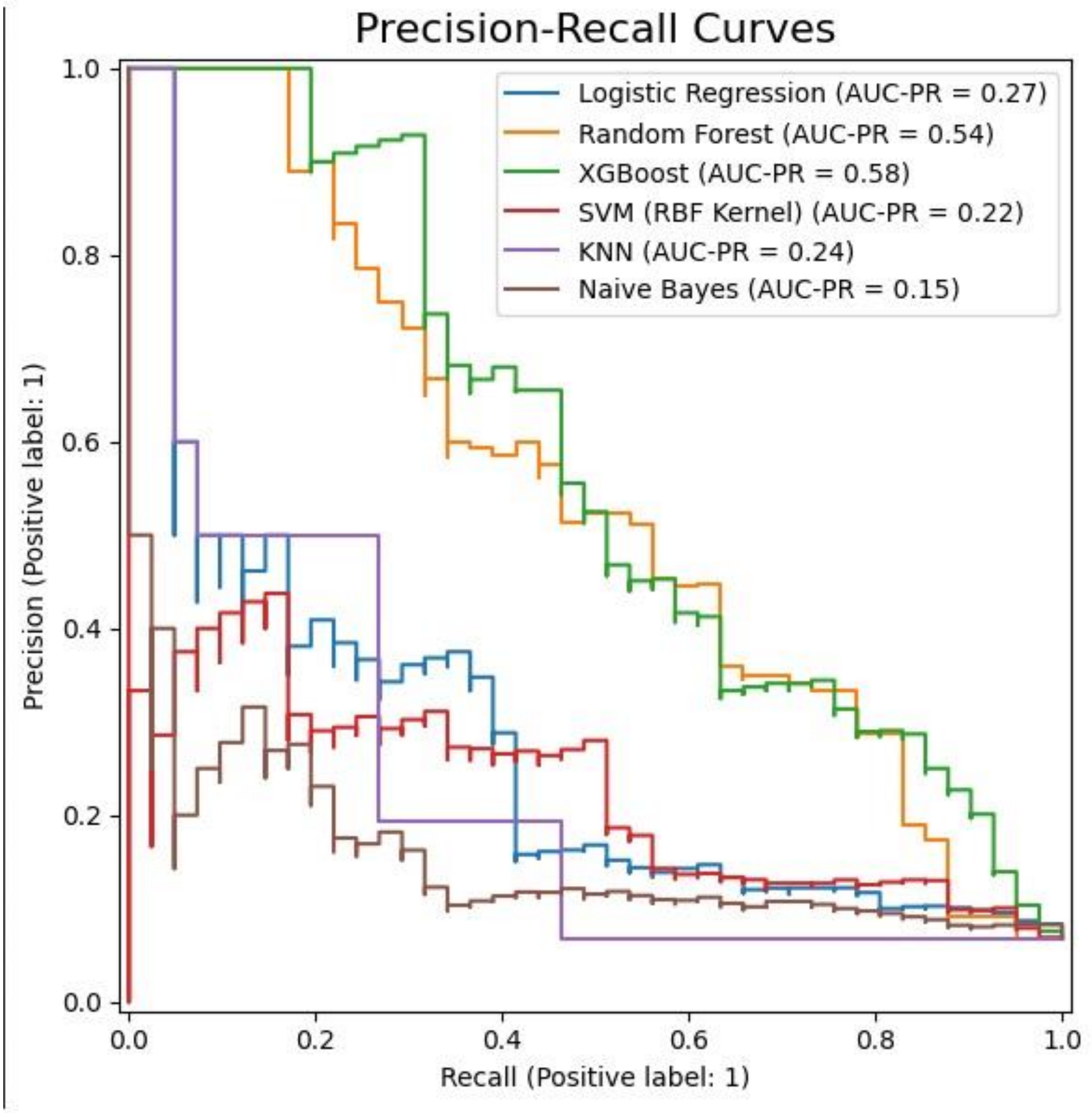

ROC Curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
Logistic Regression (ROC AUC = 0.75)
Random Forest (ROC AUC = 0.87)
XGBoost (ROC AUC = 0.90)
SVM (RBF Kernel) (ROC AUC = 0.76)
KNN (ROC AUC = 0.68)
Naive Bayes (ROC AUC = 0.68)

Figure 2



Precision-Recall Curves
Logistic Regression (AUC-PR = 0.27)
Random Forest (AUC-PR = 0.54)
XGBoost (AUC-PR = 0.58)
SVM (RBF Kernel) (AUC-PR = 0.22)
KNN (AUC-PR = 0.24)
Naive Bayes (AUC-PR = 0.15)
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

References

- Abernathy, P. M. (2018). *The expanding news desert*. University of North Carolina at Chapel Hill, Center for Innovation and Sustainability in Local Media.
- Ali, C. (2017). *Media Localism: The Policies of Place*. University of Illinois Press.
- Franklin, B. (2014). The Future of Journalism: In an age of digital media and economic uncertainty. *Journalism Studies*, *15*(5), 481–499. https://doi.org/10.1080/1461670X.2014.930254
- Hindman, M. (2018). *The internet trap: How the digital economy builds monopolies and undermines democracy*. Princeton University Press.
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
- Napoli, P. M., Stonbely, S., McCollough, K., & Renninger, B. (2018). Local journalism and the information needs of local communities: Toward a scalable assessment approach. *Journalism Practice*, 11(4), 373-395.
- Sawangarreerak, S., & Thanathamathee, P. (2020). Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression. Information, 11(11), Article 11. https://doi.org/10.3390/info11110519
- Stonbely, S. (2023). What Makes for Robust Local News Provision? Structural Correlates of Local News Coverage for an Entire U.S. State, and Mapping Local News Using a New Method. *Journalism and Media*, *4*(2), Article 2. https://doi.org/10.3390/journalmedia4020031