Feature Selection & Model Comparison Report

Executive Summary

This report presents a comprehensive analysis of feature selection techniques applied to a **phishing URL detection dataset**. The study evaluates four feature selection algorithms and compares performance across three gradient boosting models.

Dataset Overview	Value
Total Samples	235,795
Original Features	57 (numeric)
Training Set	188,636 samples (80%)
Test Set	47,159 samples (20%)
Target Classes	Legitimate (1): 134,850 / Phishing (0): 100,945

---

1. Feature Selection Algorithms

1.1 Boruta Feature Selection
> **Type:** Wrapper method using Random Forest

Metric	Value
Selected Features	52 out of 57
Execution Time	1,386.60 seconds
Iterations	100

<details>
<summary>Selected Features (52)</summary>

1. URLLength, DomainLength, URLSimilarityIndex, CharContinuationRate, TLDLegitimateProb
2. URLCharProb, TLDLength, NoOfSubDomain, NoOfObfuscatedChar, NoOfLettersInURL
3. LetterRatioInURL, NoOfDegitsInURL, DegitRatioInURL, NoOfEqualsInURL, NoOfQMarkInURL
4. NoOfAmpersandInURL, NoOfOtherSpecialCharsInURL, SpacialCharRatioInURL, IsHTTPS, LineOfCode
5. LargestLineLength, HasTitle, DomainTitleMatchScore, URLTitleMatchScore, HasFavicon
6. Robots, IsResponsive, NoOfURLRedirect, HasDescription, NoOfPopup
7. NoOfiFrame, HasExternalFormSubmit, HasSocialNet, HasSubmitButton, HasHiddenFields
8. HasPasswordField, Bank, Pay, HasCopyrightInfo, NoOfImage
9. NoOfCSS, NoOfJS, NoOfSelfRef, NoOfEmptyRef, NoOfExternalRef
10. has_no_www, num_slashes, num_hyphens, URL_Profanity_Prob, URL_NumberOf_Profanity
11. URLContent_Profanity_Prob, URLContent_NumberOf_Profanity

</details>

1.2 RFE (Recursive Feature Elimination)
> **Type:** Wrapper method using LightGBM as base estimator

| Metric | Value |

|--------|-------|
Selected Features	20 out of 57
Execution Time	53.68 seconds
Target Features	20

Selected Features (20):
#	Feature	#	Feature
1	URLLength	11	LineOfCode
2	URLSimilarityIndex	12	LargestLineLength
3	CharContinuationRate	13	HasDescription
4	TLDLegitimateProb	14	NoOfImage
5	URLCharProb	15	NoOfCSS
6	NoOfSubDomain	16	NoOfJS
7	NoOfLettersInURL	17	NoOfSelfRef
8	LetterRatioInURL	18	NoOfEmptyRef
9	SpacialCharRatioInURL	19	NoOfExternalRef
10	IsHTTPS	20	URL_Profanity_Prob

1.3 Correlation-based Feature Selection
> **Type:** Filter method (correlation threshold ≥ 0.1)

Metric	Value
Selected Features	39 out of 57
Execution Time	1.21 seconds
High Correlation Removal	>0.95 inter-feature correlation

Top 10 Correlated Features with Target:

Rank	Feature	Correlation
1	URLSimilarityIndex	0.8604
2	HasSocialNet	0.7837
3	HasCopyrightInfo	0.7428
4	HasDescription	0.6906
5	has_no_www	0.6684
6	IsHTTPS	0.6129
7	DomainTitleMatchScore	0.5835
8	HasSubmitButton	0.5790
9	IsResponsive	0.5485
10	URLTitleMatchScore	0.5384

1.4 Ensemble Feature Importance Selection
> **Type:** Aggregated importance from LightGBM, XGBoost, and CatBoost

Metric	Value
Selected Features	8 out of 57
Execution Time	5.71 seconds
Threshold	≥10% of max importance

Top 8 Features by Average Importance:

| Rank | Feature | Avg Importance |

|---|---|---|
1	URLSimilarityIndex	0.7348
2	LineOfCode	0.3397
3	LargestLineLength	0.3157
4	NoOfExternalRef	0.1865
5	URLCharProb	0.1799
6	LetterRatioInURL	0.1617
7	SpacialCharRatioInURL	0.1210
8	IsHTTPS	0.1120

---

2. Feature Selection Summary

Method	Features Selected	% of Original	Selection Time (s)
All Features	57	100.0%	0.00
Boruta	52	91.2%	1,386.60
RFE	20	35.1%	53.68
Correlation-based	39	68.4%	1.21
Ensemble Importance	8	14.0%	5.71

Common Features Across All Methods (7)

These features were consistently selected by all four methods, indicating high predictive importance:

Feature	Description
URLSimilarityIndex	Measure of URL similarity patterns
LineOfCode	Number of lines in page source
URLCharProb	Character probability in URL
LetterRatioInURL	Ratio of letters in URL
SpacialCharRatioInURL	Ratio of special characters in URL
IsHTTPS	Whether URL uses HTTPS protocol
NoOfExternalRef	Number of external references

---

3. Model Performance Results

3.1 Complete Results Table

Feature Set	Features	Model	Accuracy	Precision	Recall	F1-Score	MCC	Training Time (s)
All Features	57	LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	2.2316
All Features	57	XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	1.3402
All Features	57	CatBoost	1.0000	1.0000	1.0000	1.0000	1.0000	5.2695
Boruta	52	LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	2.0044
Boruta	52	XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	1.4674
Boruta	52	CatBoost	1.0000	1.0000	1.0000	1.0000	1.0000	7.3324
RFE	20	LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.6792
RFE	20	XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	0.7253
RFE	20	CatBoost	1.0000	1.0000	1.0000	1.0000	1.0000	5.3130

Correlation	39	LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.9181
Correlation	39	XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	1.4801
Correlation	39	CatBoost	1.0000	1.0000	1.0000	1.0000	1.0000	4.7283
Ensemble	8	LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.2035
Ensemble	8	XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	0.7310
Ensemble	8	CatBoost	1.0000	1.0000	1.0000	1.0000	1.0000	4.5433

> [!NOTE]
> All models achieved perfect accuracy (100%) across all feature sets, indicating the dataset has highly discriminative features for phishing detection.

3.2 Training Time Comparison

Feature Set	LightGBM	XGBoost	CatBoost
All Features (57)	2.23s	1.34s	5.27s
Boruta (52)	2.00s	1.47s	7.33s
RFE (20)	1.68s	0.73s	5.31s
Correlation (39)	1.92s	1.48s	4.73s
Ensemble (8)	1.20s	0.73s	4.54s

3.3 Performance vs All Features Comparison

Model	Feature Set	Feature Reduction	Accuracy Change	Time Reduction
LightGBM	Boruta	8.8%	+0.0000	10.2%
LightGBM	RFE	64.9%	+0.0000	24.8%
LightGBM	Correlation	31.6%	+0.0000	14.0%
LightGBM	**Ensemble**	**86.0%**	+0.0000	**46.1%**
XGBoost	Boruta	8.8%	+0.0000	-9.5%
XGBoost	RFE	64.9%	+0.0000	45.9%
XGBoost	Correlation	31.6%	+0.0000	-10.4%
XGBoost	**Ensemble**	**86.0%**	+0.0000	**45.5%**
CatBoost	Boruta	8.8%	+0.0000	-39.1%
CatBoost	RFE	64.9%	+0.0000	-0.8%
CatBoost	Correlation	31.6%	+0.0000	10.3%
CatBoost	**Ensemble**	**86.0%**	+0.0000	**13.8%**

---

4. Best Configurations

Category	Best Configuration
Best Accuracy	All models achieved 1.0 (perfect)
Best F1-Score	All models achieved 1.0 (perfect)
Best MCC	All models achieved 1.0 (perfect)
Best Efficiency (Accuracy/Time)	**XGBoost + RFE (20 features)** - 1.0 accuracy in 0.73s

---

5. Key Findings & Recommendations

Key Findings

Finding	Details
Perfect Classification	All models achieved 100% accuracy regardless of feature selection method
Most Efficient Feature Set	Ensemble Importance with only **8 features** maintains perfect accuracy
Fastest Model	XGBoost consistently showed lowest training times
Most Important Feature	**URLSimilarityIndex** (0.86 correlation with target)

Recommendations

> [!IMPORTANT]
> **For Production Deployment:**
> - Use **Ensemble Feature Selection (8 features)** to minimize model complexity
> - Choose **XGBoost** for fastest inference times
> - This achieves **86% feature reduction** with no accuracy loss

> [!TIP]
> **For Research/Experimentation:**
> - Use **RFE (20 features)** for a balance between dimensionality and interpretability
> - Consider the 7 common features as core predictive signals
> - **URLSimilarityIndex** alone provides 86% correlation with the target

Feature Selection Trade-offs

```
┌──────────────────────────────────────────────────────────┐
│  Feature Selection Methods Comparison                    │
├──────────────────────────────────────────────────────────┤
│                                                          │
│  Boruta (52 features)                                    │
│  ████████████████████████████████████████████████       │
│  Most comprehensive - 91% of features retained           │
│                                                          │
│  Correlation (39 features)                               │
│  ████████████████████████████████████                   │
│  Fast selection - Best for quick analysis                │
│                                                          │
│  RFE (20 features)                                       │
│  ██████████████████                                      │
│  Balanced - Good for model interpretability              │
│                                                          │
│  Ensemble (8 features)                                   │
│  ████████                                                │
│  Most aggressive - Best for production efficiency        │
│                                                          │
└──────────────────────────────────────────────────────────┘
```

---

6. Conclusion

This analysis demonstrates that **effective feature selection can dramatically reduce model complexity** (by up to 86%) **without sacrificing predictive performance** on this phishing URL detection task.

The **7 consistently important features** identified across all methods represent the core signals for distinguishing legitimate from phishing URLs:

1. **URLSimilarityIndex** - Most discriminative feature
2. **LineOfCode** - Page content complexity
3. **URLCharProb** - URL character patterns
4. **LetterRatioInURL** - URL composition
5. **SpacialCharRatioInURL** - Special character usage
6. **IsHTTPS** - Security protocol
7. **NoOfExternalRef** - External resource usage

> [!CAUTION]
> The perfect accuracy (100%) achieved by all models may indicate potential **data leakage** or an overly simplistic classification task. Further validation with cross-validation and external test sets is recommended.

Report generated from `feature-selection.ipynb` analysis