Feature Importance Analysis Report

1. Executive Summary

This report analyzes JavaScript API usage patterns to identify features most predictive of different tracking behaviors. Using machine learning (Random Forest classification with TF-IDF vectorization), we ranked features by their discriminative power.

Dataset Overview

Metric	Value
Total Websites Analyzed	49
Unique Features Detected	1,000
Categories Identified	2

Category Distribution

Category	Count	Percentage	Distribution
Advertising	35	71.4%	
Analytics	14	28.6%	

2. Key Findings

Top Feature Categories in Tracking Detection

The most predictive features fall into these categories:

- **DOM Properties**: 3 features in top 20
- **Network**: 2 features in top 20
- **Page Lifecycle**: 2 features in top 20
- **DOM Traversal**: 2 features in top 20
- **CSS Styling**: 1 feature in top 20
- **Event Handling**: 1 feature in top 20
- **Timers**: 1 feature in top 20
- **Observers**: 1 feature in top 20
- **Performance**: 1 feature in top 20
- **URL/Location**: 1 feature in top 20
- **HTML Elements**: 1 feature in top 20
- **Fingerprinting**: 1 feature in top 20
- **Error Monitoring**: 1 feature in top 20
- **Script/Document**: 1 feature in top 20
- **Frameworks**: 1 feature in top 20

3. Top 20 Most Predictive Features

These features have the highest discriminative power for classifying tracking behavior:

Rank	Feature	Importance	Relative	Category
1	`display`	0.0247		CSS Styling
2	`customevent`	0.0205		Event Handling
3	`setinterval`	0.0138		Timers
4	`nexthopprotocol`	0.0137		Network
5	`intersectionratio`	0.0134		Observers
6	`id`	0.0117		DOM Properties
7	`nodename`	0.0110		DOM Properties
8	`visibilitystate`	0.0104		Page Lifecycle
9	`getentriesbytype`	0.0099		Performance
10	`hostname`	0.0094		URL/Location
11	`abortcontroller`	0.0093		Network
12	`htmlulistelement`	0.0090		HTML Elements
13	`font`	0.0080		Fingerprinting
14	`queryselector`	0.0079		DOM Traversal
15	`parentnode`	0.0078		DOM Traversal
16	`__sentry__`	0.0077		Error Monitoring
17	`namednodemap`	0.0076		DOM Properties
18	`currentscript`	0.0076		Script/Document
19	`prerendering`	0.0076		Page Lifecycle
20	`__reactfiber`	0.0072		Frameworks
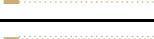
Note: The prominence of timing/observation APIs (`setinterval`, `intersectionratio`, `visibilitystate`) suggests that tracking scripts heavily rely on monitoring user behavior and page state changes.

---

4. Category-Specific Feature Analysis

Below are the most characteristic features for each tracking category, ranked by TF-IDF score within that category.

4.1 Advertising

Scripts focused on ad delivery, targeting, and conversion tracking. Often communicate with ad networks and DSPs.

Rank	Feature	TF-IDF Score	Relevance
1	window	0.4279	
2	htmldivelement	0.2224	
3	htmldocument	0.1329	
4	htmlanchorelement	0.0932	
5	setattribute	0.0841	
6	getattribute	0.0830	
7	nodetype	0.0816	
8	htmlspanelement	0.0728	
9	addeventlistener	0.0635	
10	cssstyledeclaration	0.0614	

> **Notable:** Top features include window, htmldivelement, htmldocument

4.2 Analytics

Scripts that measure user engagement, page performance, and behavioral patterns for business intelligence.
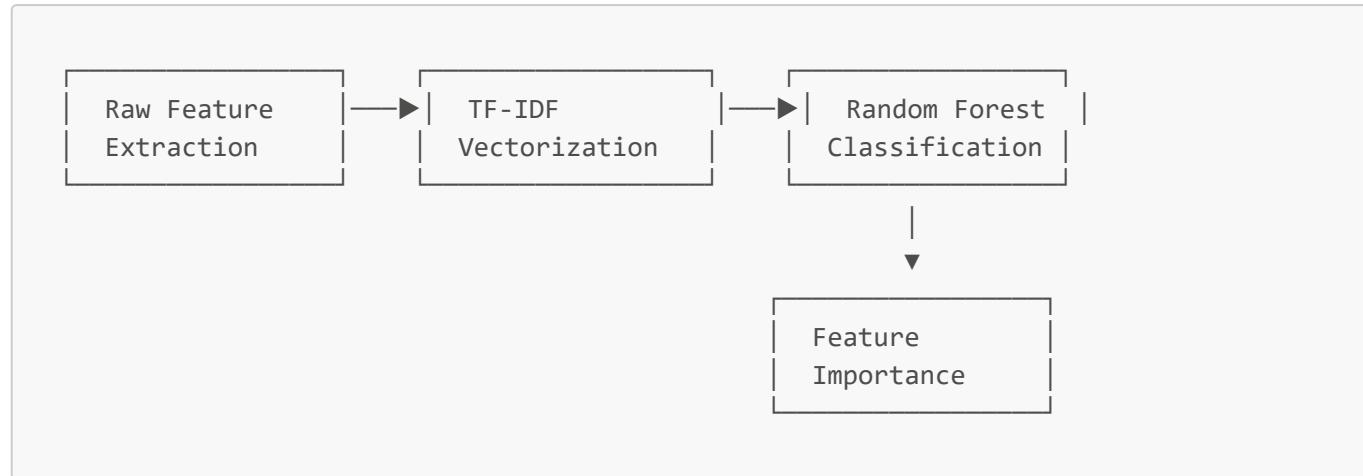
Rank	Feature	TF-IDF Score	Relevance
1	window	0.3257	
2	htmldivelement	0.2293	
3	htmldocument	0.0979	
4	htmlspanelement	0.0874	
5	addeventlistener	0.0798	
6	setattribute	0.0773	
7	nodetype	0.0766	
8	__reactprops	0.0764	
9	performance	0.0757	
10	getattribute	0.0753	

> **Notable:** Top features include window, htmldivelement, htmldocument

5. Methodology

Analysis Pipeline

```
┌─────────────┐    ┌─────────────┐    ┌─────────────┐
│ Raw Feature │───▶│   TF-IDF    │───▶│Random Forest│
│ Extraction  │    │Vectorization│    │Classification│
└─────────────┘    └─────────────┘    └─────────────┘
                                             │
                                             ▼
                                     ┌─────────────┐
                                     │   Feature   │
                                     │ Importance  │
                                     └─────────────┘
```

Technical Details

Component	Configuration
Vectorizer	TF-IDF with max 1,000 features
Classifier	Random Forest (100 estimators)
Feature Weighting	Usage count (capped at 100)
Ranking Method	Gini importance from RF

6. Appendix: Feature Category Definitions

Category	Description	Example Features
Fingerprinting	Device/browser identification techniques	canvas, webgl, audiocontext, timezone
Storage	Client-side data persistence	cookie, localstorage, sessionstorage, indexeddb
Network	HTTP requests and data transmission	fetch, xmlhttprequest, beacon, websocket
DOM Traversal	Finding elements in page structure	queryselector, getelementby, parentnode, childnodes
DOM Modification	Changing page content dynamically	createelement, appendchild, removechild, innerhtml
DOM Properties	Reading element attributes	nodetype, nodename, tagname, namednodemap
HTML Elements	Specific HTML element types	htmldiv, htmlspan, htmlanchor, htmlinput

Category	Description	Example Features
Event Handling	User interaction listeners	addeventlistener, removeeventlistener, customevent, dispatchevent
Timers	Scheduled code execution	setinterval, settimeout, requestanimationframe, cancelanimationframe
Observers	Monitoring page/element changes	intersectionobserver, intersectionratio, mutationobserver, resizeobserver
Page Lifecycle	Page state and visibility	visibilitystate, hidden, prerendering, readystate
Performance	Timing and metrics collection	performance, getentriesbytype, getentriesbyname, timing
URL/Location	Page address information	hostname, href, origin, pathname
Navigator	Browser/device information	navigator, useragent, language, languages
CSS Styling	Visual appearance control	style, display, cssstyledeclaration, getcomputedstyle
Layout/Geometry	Element size and position	getboundingclientrect, offsetwidth, offsetheight, clientwidth
User Input	Keyboard, mouse, touch events	scroll, click, mouse, touch
Pointer Events	Mouse/touch tracking	pointerdown, pointerup, pointermove, pointerenter
Script/Document	Core document APIs	currentscript, script, document, window
Error Monitoring	Error tracking services	sentry, bugsnag, rollbar, errorhandler
Frameworks	JS framework internals	react, angular, vue, jquery