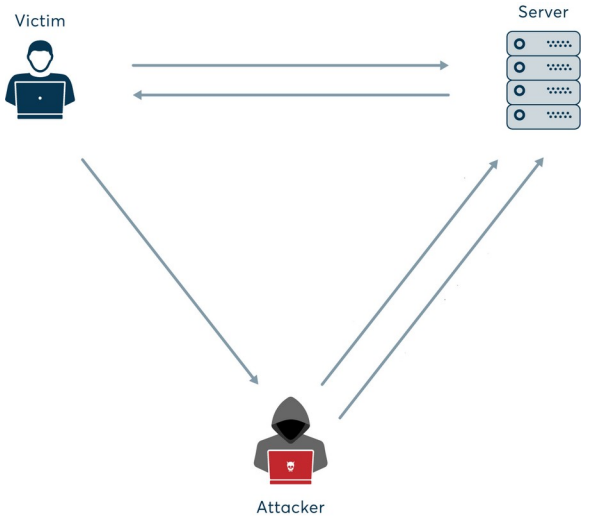
Web User Identification
Using
Sequential Pattern Mining

Goal: Intruder Detection

Distinguish between Alice's browsing patterns and those of unauthorized individuals (e.g., hackers)



Overview

- **Source:** Browsing data from Blaise Pascal University servers (Kaggle Competition **“Catch Me If You Can”**)

Overview

- **Source:** Browsing data from Blaise Pascal University servers (Kaggle Competition **“Catch Me If You Can”**)

- **Methodology:** Sequential Pattern Mining **(SPM)**

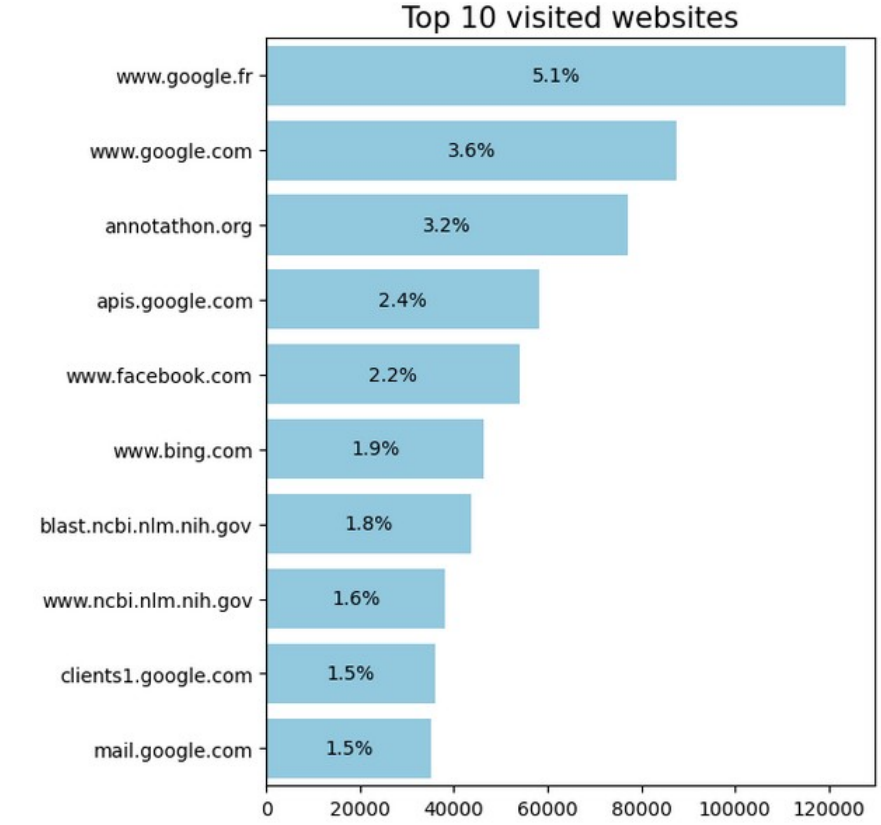
Overview

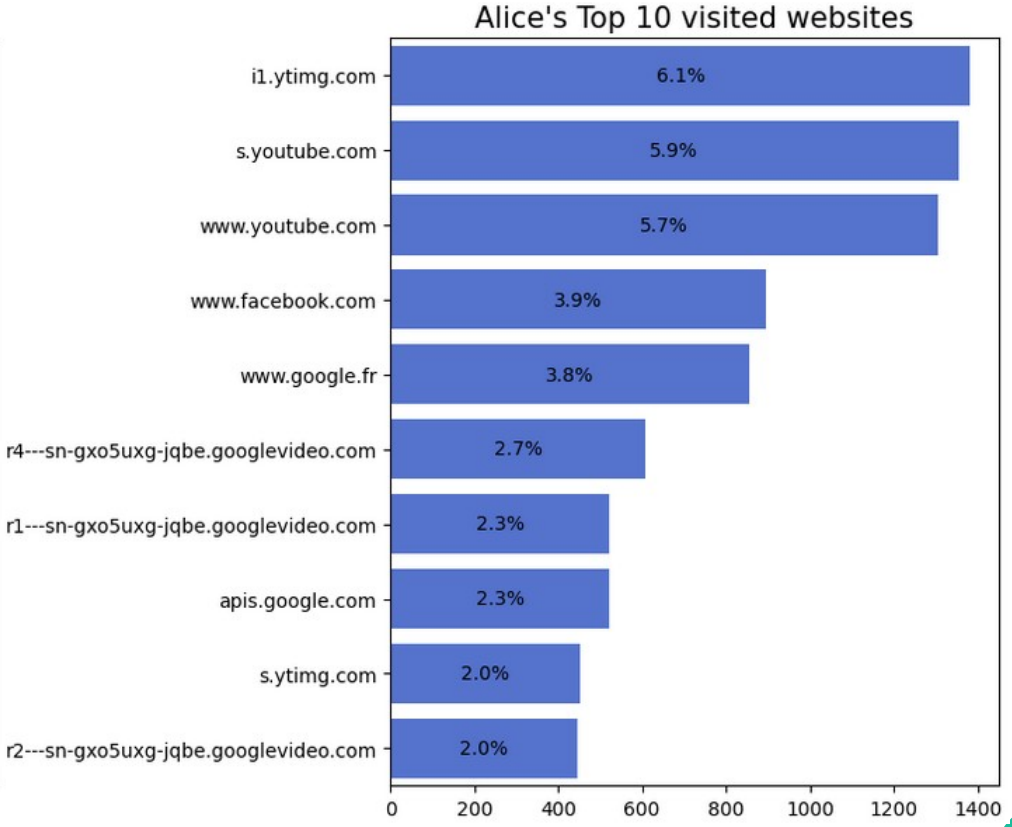
- **Source:** Browsing data from Blaise Pascal University servers (Kaggle Competition **“Catch Me If You Can”**)

- **Methodology:** Sequential Pattern Mining **(SPM)**

- **Training Dataset:**
  - 250k rows (Max 30 min long session OR max 10 websites)
  - 20 Features (Webpages and Timestamps)

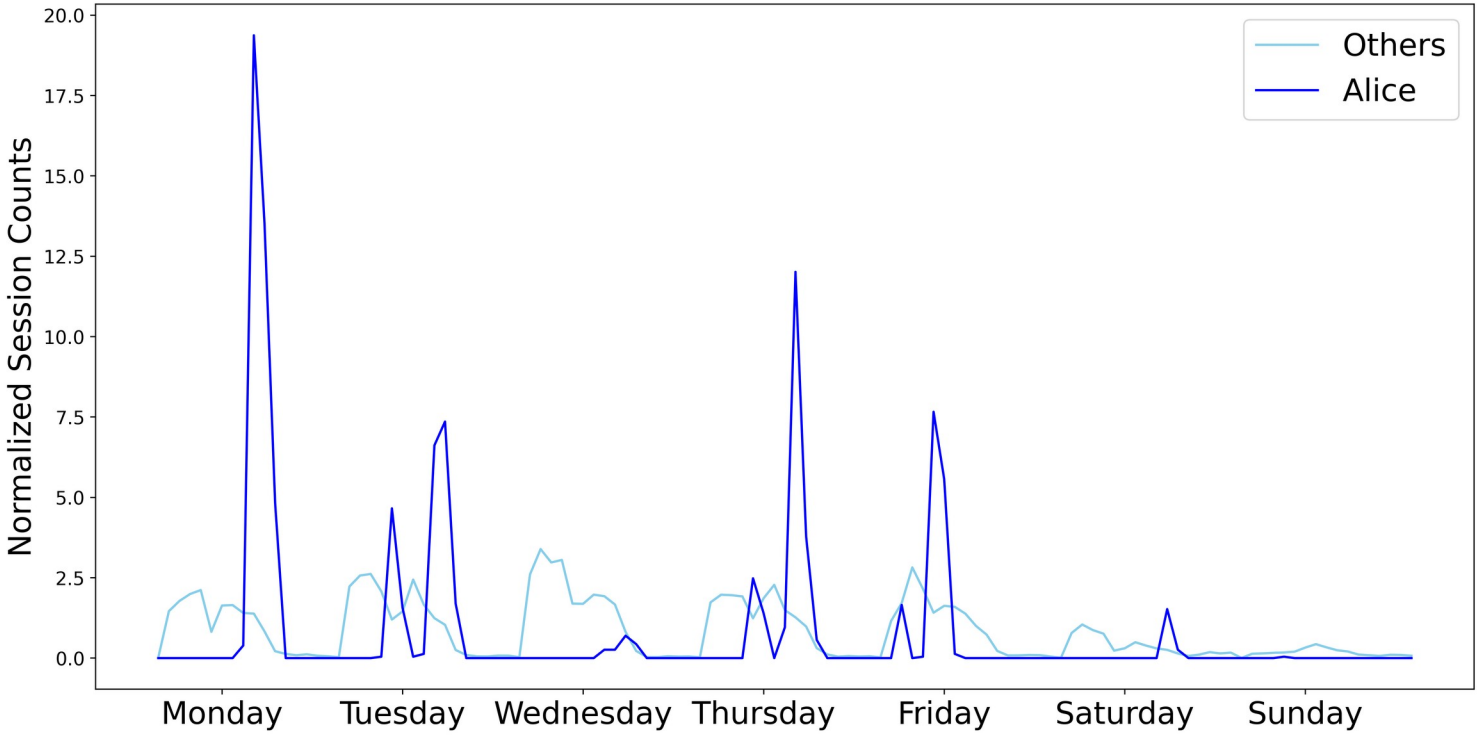
Top 10 of most visited websites





More insights from time sessions

Time range:
Jan 2013 –
Apr 2014

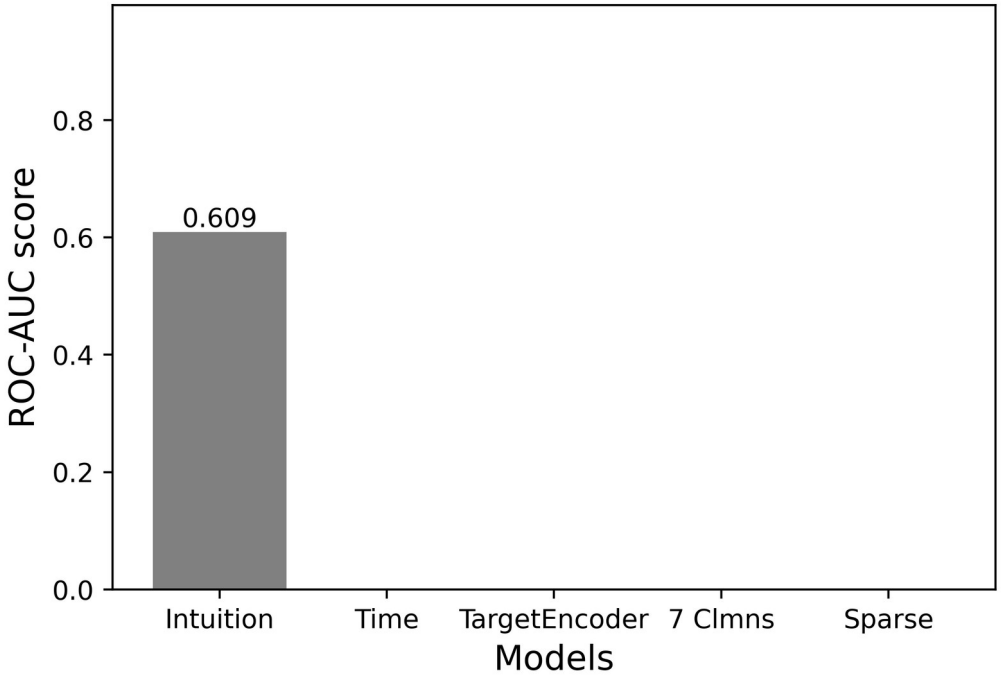


Potential Feature as Alice Seems to Have Few Preferred Days/Hours

Results

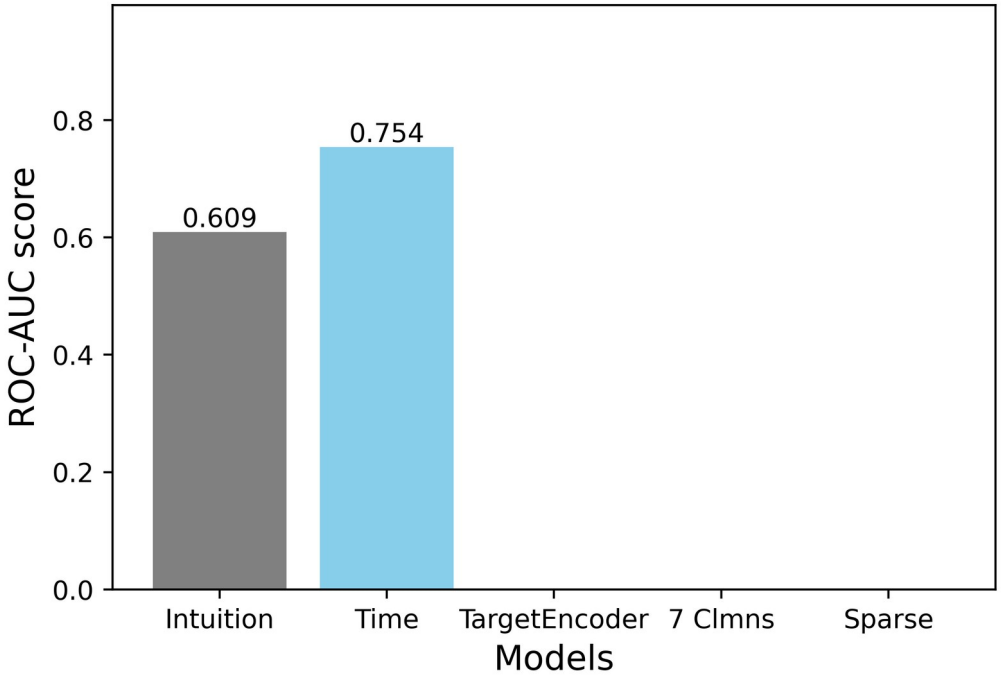
Results

- **Intuition Model (OHE)**



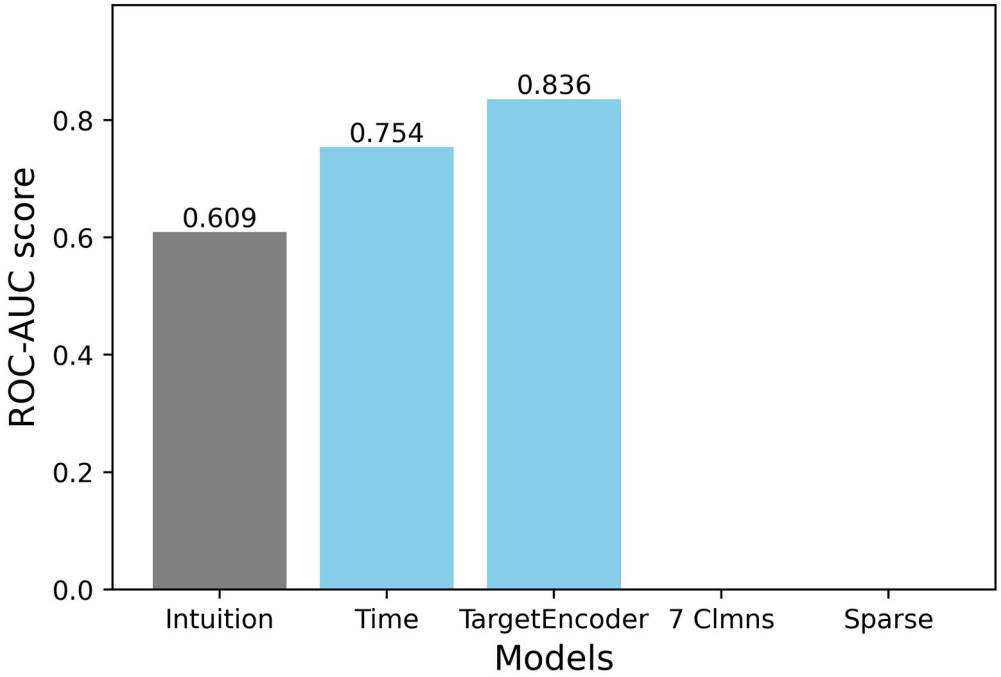
Results

- **Intuition Model (OHE)**
- **+ Time (OHE)**



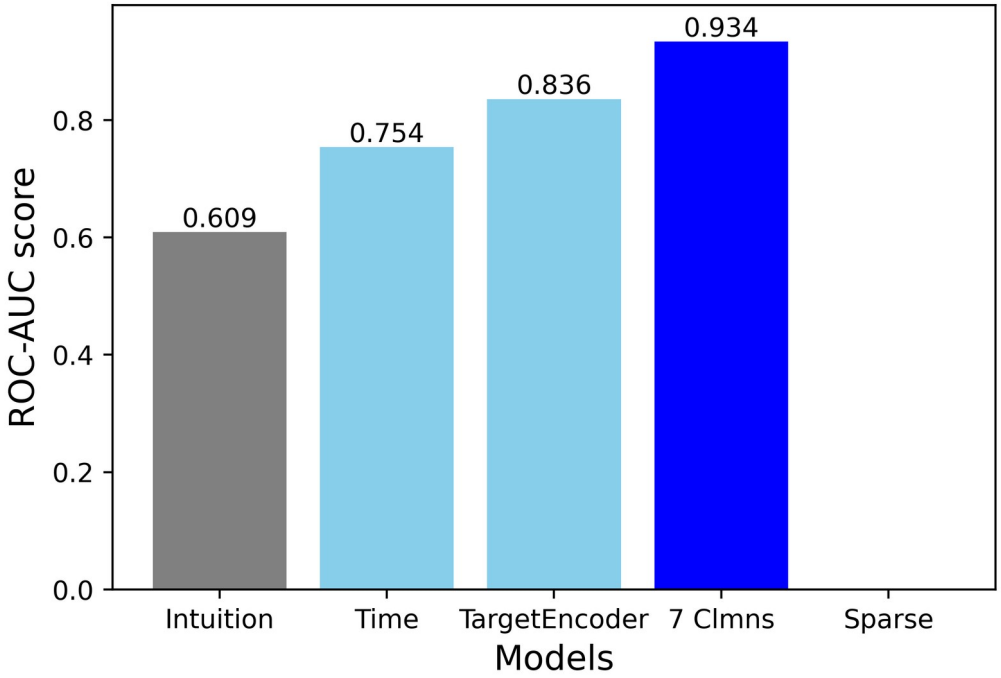
Results

- Intuition Model (OHE)
- + Time (OHE)
- + TargetEncoder



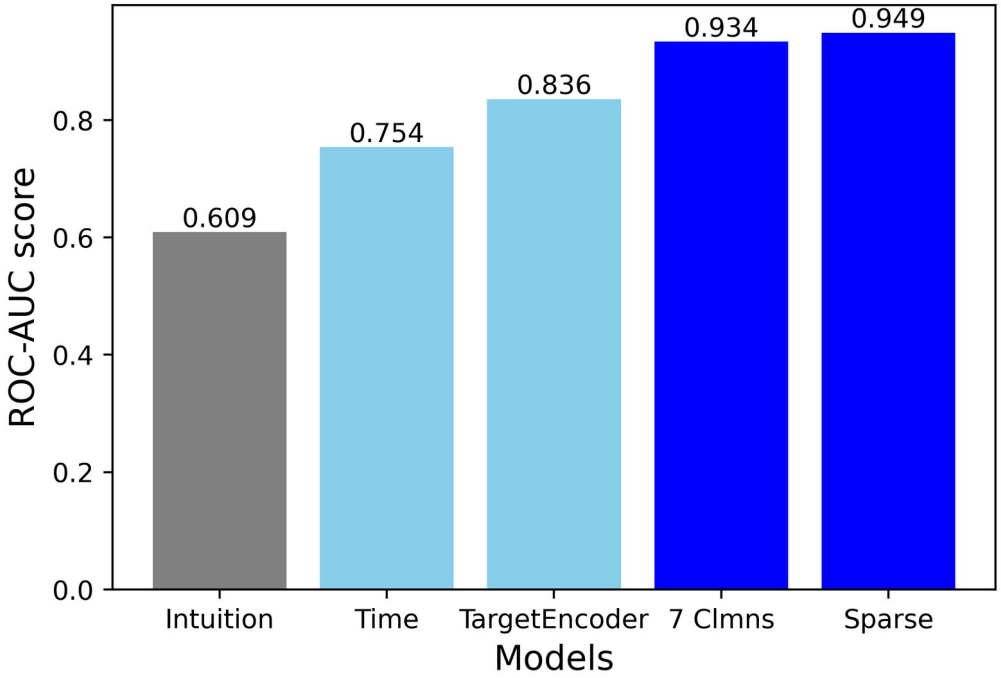
Results

- Intuition Model (OHE)
- + Time (OHE)
- + TargetEncoder
- **Time Features → 7**



Results

- Intuition Model (OHE)
- + Time (OHE)
- + TargetEncoder
- **Time Features → 7**
- **Websites→Sparse Matr.**



Kaggle Submission

- 5415 Participants
- 66361 Submissions

- Intuition model: 0.61 → Final model: 0.95

Thank you