IPSW 2023 Radio-Canada Final Presentation

Determining the right moment for suggesting the creation of an account

Mario Canche (CIMAT), Andrea Ek (CIMAT), Michael R Lindstrom (UTRGV), Corentin Lonjarret (Radio-Canada), Patrick Mesana (HEC Montreal), Carlos Montes (UTRGV), Nicolas Schönau (Radio-Canada), Omar Sharif (UTRGV), Marziyeh Talebian (Concordia), Louis Willems (Radio-Canada)

August 25, 2023

Agenda

Application Problem Statement

What is the problem?

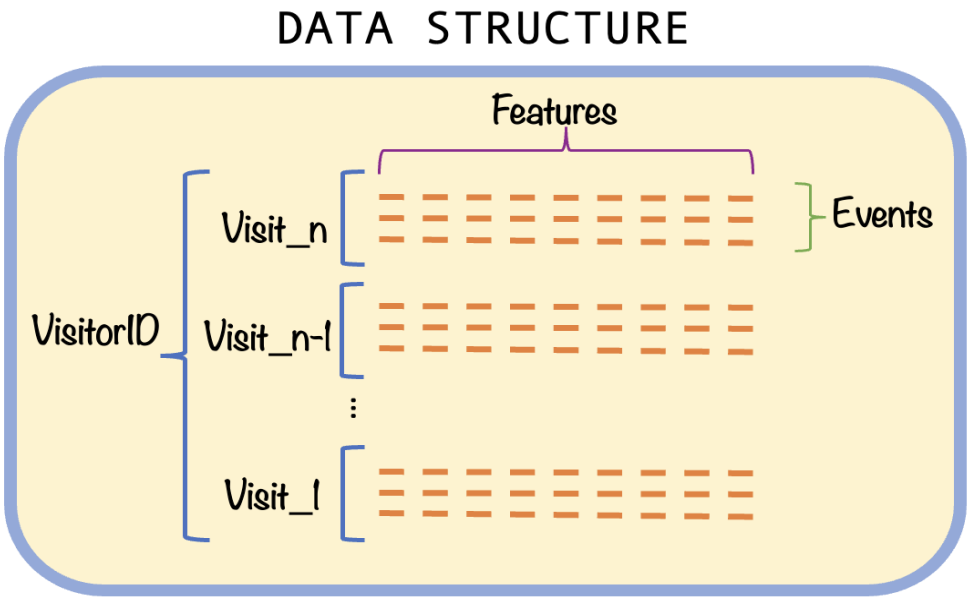
1. CBC/Radio-Canada supports Canadian culture and democratic life with a diversity of content
2. A strategic aim of the website is to encourage visitors to create accounts to enable personalized experiences
3. Users interact with the platform in manifold ways and very few users create accounts
4. Radio-Canada wants to **determine the optimal moment to suggest account creation to users** while also **mitigating the annoyance of such a prompt**

Mathematical Problem Statement

Where is the math?

1. We seek to identify **sequences of events** that are closely tied with account creation
2. We seek to estimate the **probability a user on their journey will create an account**
3. We seek to identify **similarities (clustered features)** between the users who create accounts; likewise for those who do not

Data



How data are handled?

1. data are a collection of webpage interactions with timestamps from Apr-May 2023, including the type of activity, and unique ID per visitor
2. some filtering was needed, e.g., filtering out events after account creation or users with presence before sample
3. some users may delete their cookies and appear under different IDs, bots may be in the dataset, etc.

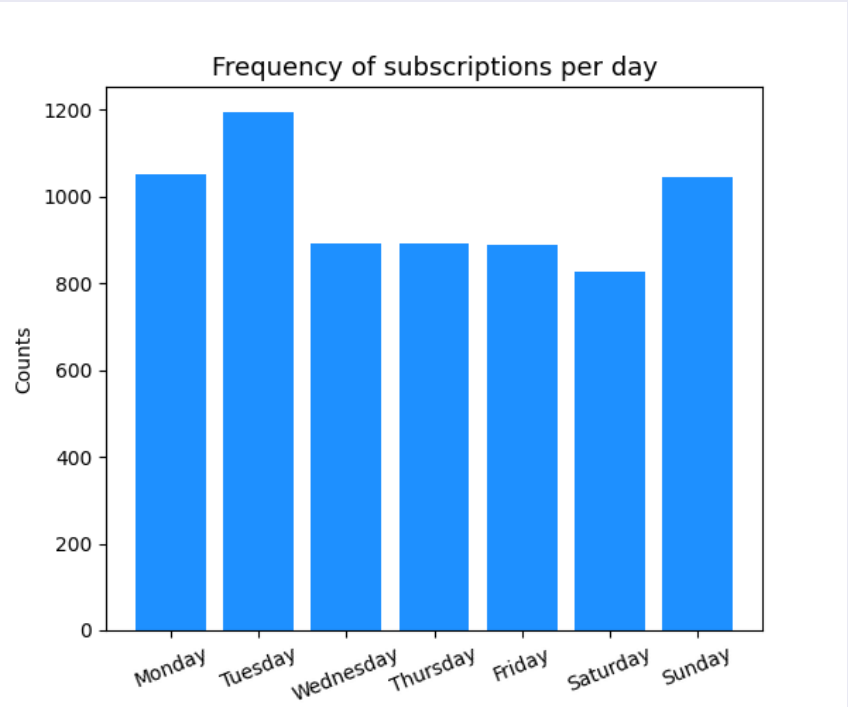
Methods

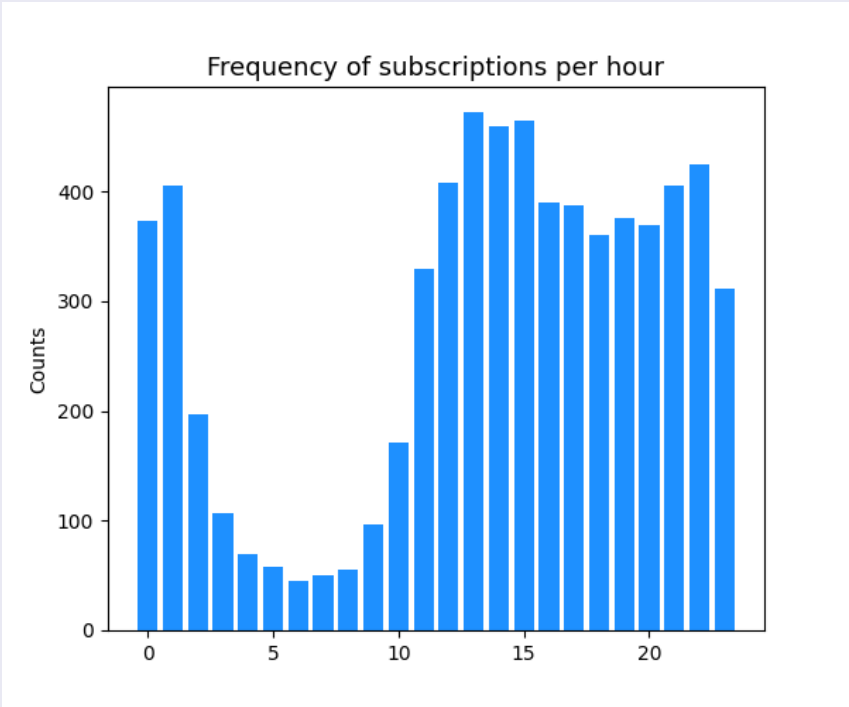
Summary of our methods

1. Visualizations and Clustering: clustering visitor activity to understand user groups and identifying features that distinguish account-creators
2. Interaction Driven Creation Model: modelling the probability a visitor will create an account given their current webpage interaction status
3. Events Embedding Analysis: representing events as vectors to find correlated events with account creation events

Visualizations and Clustering

Temporal variations in account creation



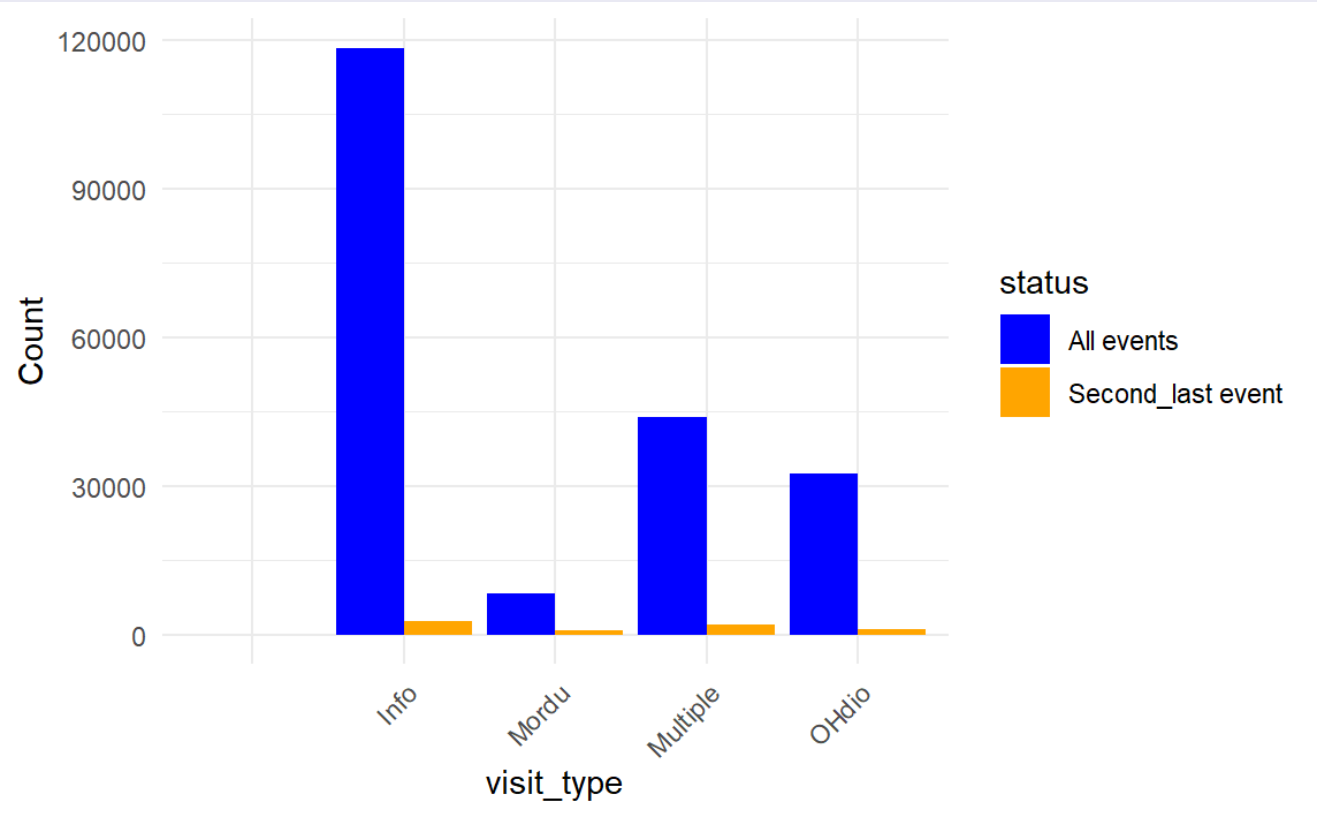


Visualizations and Clustering

Most recent visit types by different user groups

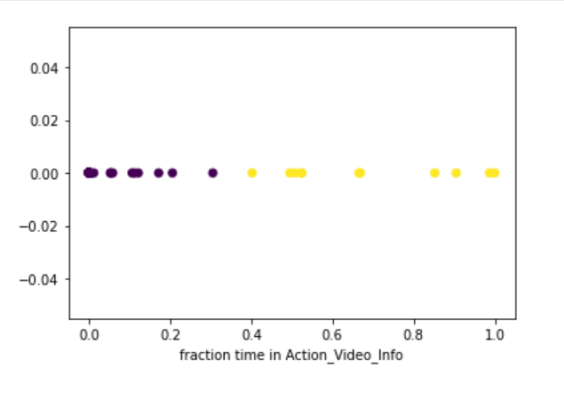
Visualizations and Clustering

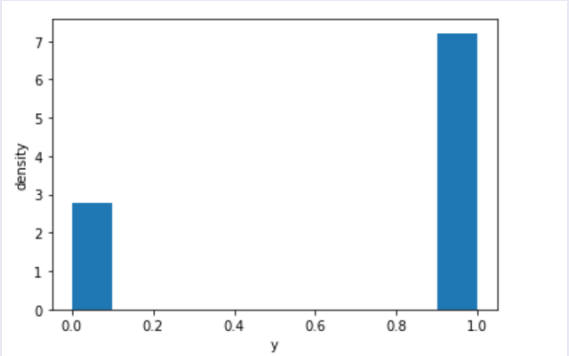
Final visit types recorded before account creation

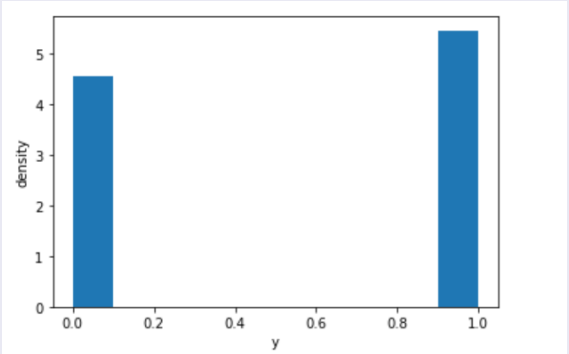


Visualizations and Clustering

Kmeans analysis







Interaction Driven Creation Model

Assumptions and derivation

1. Let there be U users with data
2. Given the current user interaction $x \in \mathbb{R}^k$, let $Y_x \in \{0, 1\}$ be a r.v. for a subsequent account creation (0=no, 1=yes)
3. Model $\Pr(Y_x = 0|x) = \sigma(\theta^T \hat{x})$ where $\theta \in \mathbb{R}^{k+1}$ is a parameter, $\hat{x} = (1, x^{(1)}, ..., x^{(k)})^T$, and $\sigma : z \mapsto \frac{1}{1+\exp(-z)}$ is the sigmoid function
4. To user i, let N_i be either the number of interactions leading up to but not including account creation (if they do) or their number of interactions otherwise
5. Let $x_{ij} \in \mathbb{R}^k$ be the jth interaction of user i
6. The account dataset is $\mathcal{D}|_x = \{\{y_{ij}\}_{j=1}^{N_i}\}_{i=1}^{U}$, a collection of creation statuses with y_{ij} the realization of $Y_{x_{ij}}$

Interaction Driven Creation Model

Assumptions and derivation

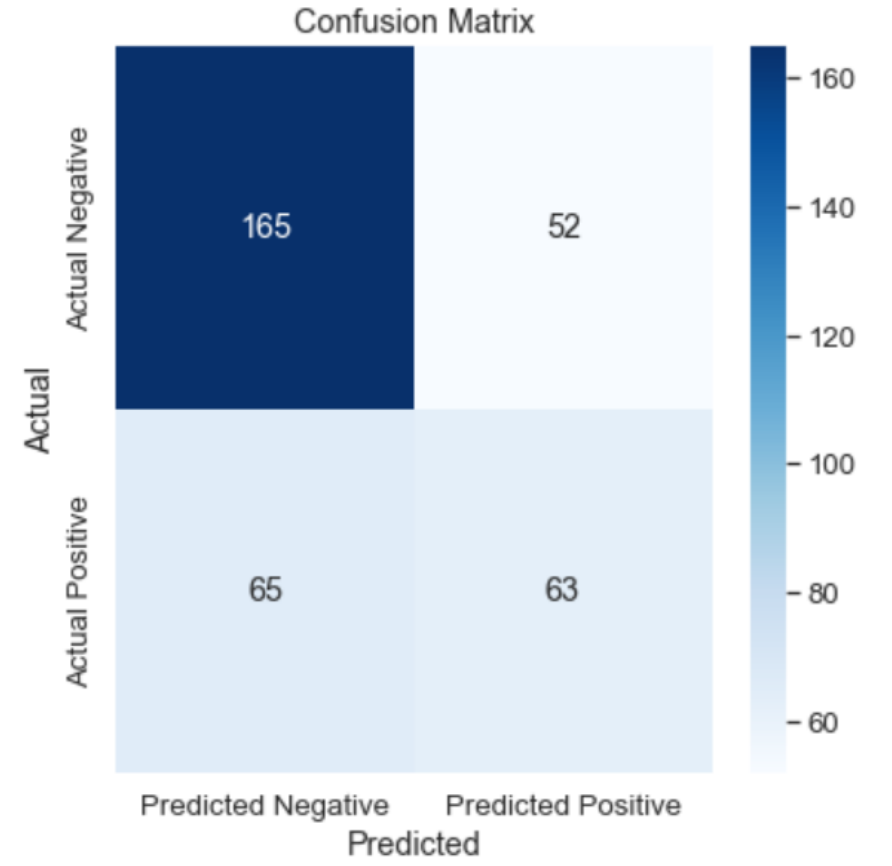
1. **Assume (approximate) decoupling** to generate separate experiments – effectively a **Logistic Regression**
2. The log-likelihood is then

$$\mathcal{L}(\mathcal{D}|_x|\theta) = \sum_{i=1}^{U} \mathbb{1}_{y_{i,N_i}=1} \left(\sum_{j=1}^{N_i-1} \log \sigma(\theta^T \hat{x}_{ij}) + \log(1 - \sigma(\theta^T \hat{x}_{i,N_i})) \right) + \sum_{i=1}^{U} \mathbb{1}_{y_{i,N_i}=0} \left(\sum_{j=1}^{N_i} \log \sigma(\theta^T \hat{x}_{ij}) \right)$$

   where $\mathbb{1}$ is the indicator function
3. This can be maximized with **gradient ascent**

Interaction Driven Creation Model

Results



	precision	recall	f1-score	support
0	0.72	0.76	0.74	217
1	0.55	0.49	0.52	128
accuracy			0.66	345
macro avg	0.63	0.63	0.63	345
weighted avg	0.65	0.66	0.66	345

Interaction Driven Creation Model

Results

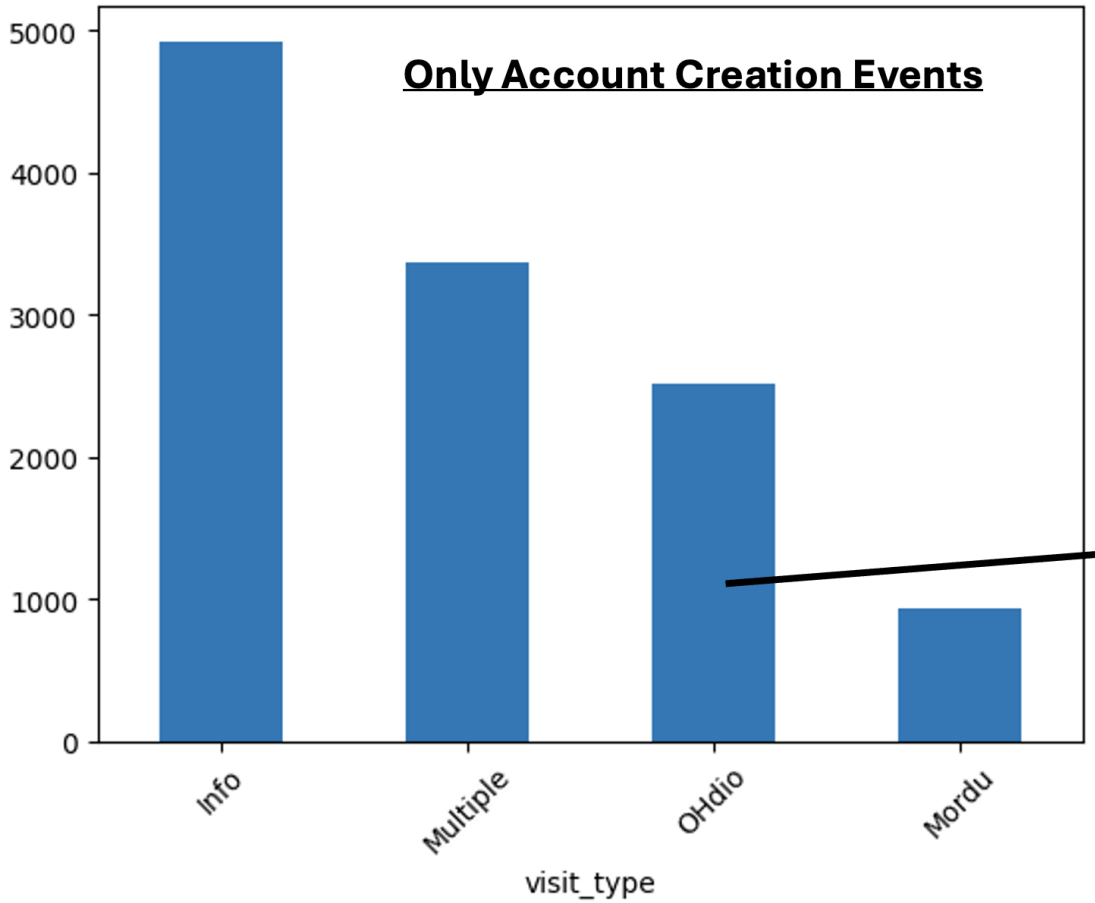
1. Looking at θ gives insight into user account creation patterns and probabilities of account creation

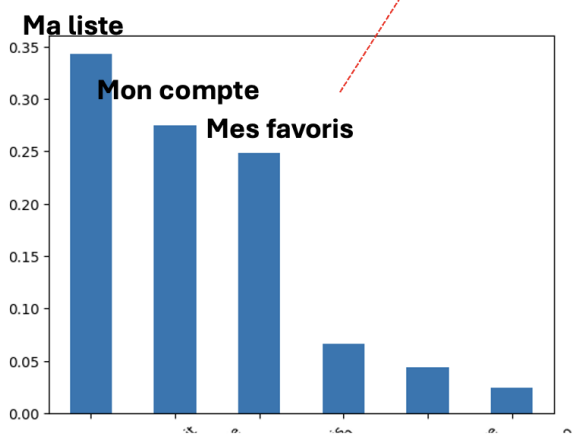
θ component	value
intercept	-0.022
duration	0.003
day_of_visit_4	0.030
day_of_visit_1	-0.048
cookies_N	-0.044
detailed_event_Écoute_OHdio	-0.039

Table: Positive coefficients make a user less likely to create an account, negative coefficients make them more likely to create an account.

Events Embedding Analysis

What is the immediate context of an account creation event?



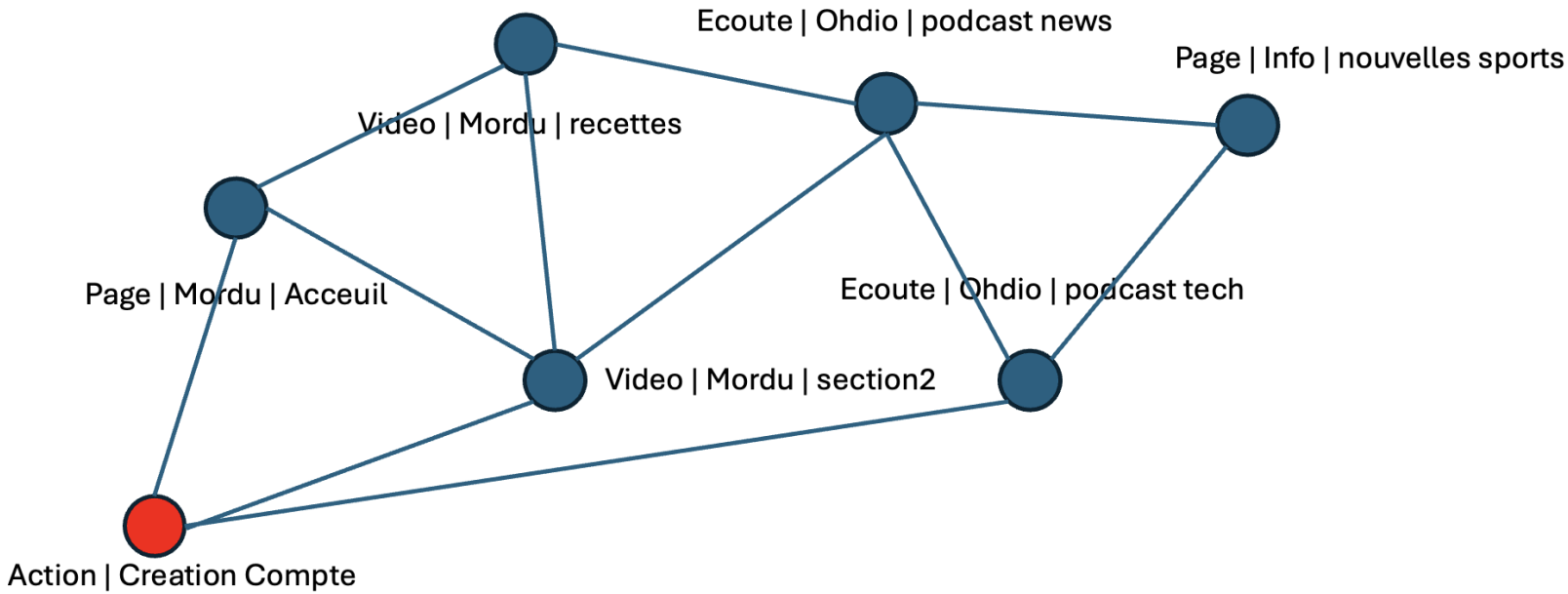


The infamous "102" column

Events Embedding Analysis

How to represent events and their context?

Events as a graph



Events Embedding Analysis

Skip-Gram with Negative Sampling (Word2Vec)

Estimate the prob of being in the same context

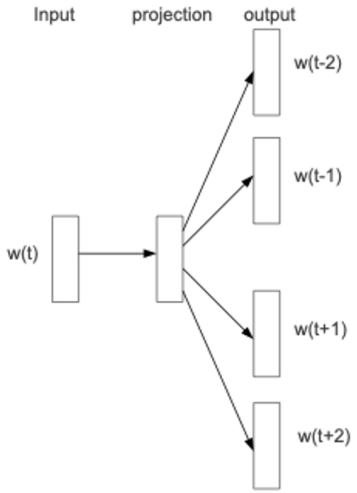
$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^\top v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^\top v_{w_I}\right)}$$

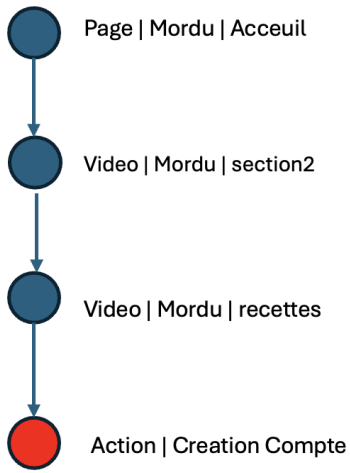
Negative Sampling trick - predicting if pairs are in the same context

max $\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N)\right]$

$P_n(w) = U(w)^{0.75}/Z$

$P(D = 0|w, c) = 1 - P(D = 1|w, c)$





Events Embedding Analysis

"Vocabulary" has 553 different events

	post_page_event	detailed_event	visit_type	post_prop5
0	0	Page_Info	Info	abitibitemiscamingue
1	0	Page_Info	Info	acadie
2	0	Page_Info	Info	alberta
3	0	Page_Info	Info	alimentation
4	0	Page_Info	Info	ar
...	...	...	...	...
548	100	Action_creation_compte	OHdio	no_post_prop5_info
549	100	Action_creation_compte	no_visit_type_info	no_post_prop5_info
550	101	Action_autre	Info	no_post_prop5_info
551	102	Action_autre	Info	no_post_prop5_info
552	102	Action_autre	Mordu	recettes

Events Embedding Analysis

DEMO

Interpretations and Insights

Main Findings

1. Sunday, Monday and Tuesday are the days where more accounts have been created
2. Visitors who experienced multiple events show a higher rate of account creation
3. Probability of account creation can be modelled, suggesting: Monday is a better day for account creation, Écoute_OHdio is a good event for account creation, and that users who have longer events may be less likely to create an account
4. Correlation between creation action from a section with events from the same section

Future Work

More to explore...

1. Use LSTM recurrent networks to model the sequential problem with the numerical variables.
2. Interaction Driven Creation Model could perhaps be improved in performance with a better model/handling of features
3. Add more features, e.g., navigation levels, to increase number of events
4. Encode sequences as context to have sequence embedding
5. Strengthening and identifying the common interpretations between the complementary models

Acknowledgments

We would like to thank...

1. Radio Canada and their representatives for contributing this problem and their availability throughout the workshop in resolving our confusions
2. Odile Marcotte and Nancy Laramée for their organization
3. CRM and IVADO for hosting the event