

Геоаналитика.

Куда можно копать

2024.03.21

DataFusion Contest 2024

Задача

- 8157 локаций. В 1657 есть банкоматы. В 8154 совершались оплаты.
- Даны транзакции клиента (кроме снятия наличных)
- Нужно для каждой из 1657 локаций предсказать вероятность снятия наличных клиентом

	h3_09	customer_id	datetime_id	count	sum	avg	min	max	std	count_distinct	mcc_code
0	8911aa4c62fffff	1	3	1	3346.65	3346.650	3346.65	3346.65	NaN	1	13
1	8911aa7b5b3ffff	4	3	1	450.00	450.000	450.00	450.00	NaN	1	8
2	8911aa63623ffff	5	3	10	11035.69	1103.569	59.00	3620.18	1190.530333	6	13
3	8911aa48577ffff	9	2	2	628.00	314.000	295.00	333.00	26.870058	2	5
4	8911aa78297ffff	11	2	1	4155.00	4155.000	4155.00	4155.00	NaN	1	10

Нужны новые идеи

- Упростить задачу
- Разбить на подзадачи
- Новые признаки
- Уточнить данные
- Extreme Multi-label
- VW BFG9000
- Кликстрим
- Автоэнкодеры



Упростить задачу

- Нужно предсказать снятие для 1657 банкоматов
- 2^{1657} вариантов
- Можно ли сделать задачу проще?
- В обучении 10571 уникальная комбинация
- Это примерно 2^{14}

Кодирование таргета

- Можно построить преобразование в 14 бинарных признаков
- Механически — перенумеруем, возьмем биты
- Получим 14 плотных бинарных таргетов
- Удобные для обучения
- Можно факторизовать, получим небинарный таргет для регрессии

Разбить на подзадачи

- Зачем люди снимают деньги?
 - Бабушки не доверяют банку
 - Рынок, чаевые
 - Денег детям/родителям/в школу
 - Теневая экономика
 - Скрыть транзакцию из выписки
 - Разовые крупные сделки / перенос между счетами

Как люди перемещаются

- Дом-дорога-работа-магазин
- Дом-дорога-офис-ресторан
- Дом-банкомат-дом

Но можно

- Расшифровки МСС нет, разделить трудно
- Еда — часто, понемногу?
- Кутим — редко, крупно, по ночам?

- Обучить несколько моделей, смесь экспертов.
- МоЕ - не только про LLM

Можем ли мы знать?

- Где живет клиент
 - Мелкие расходы утром и вечером, много терминалов
- Где работает человек
 - Мелкие расходы днем, много
- Где закупается
 - Большие расходы днем, много, много терминалов
- Где кутит
 - Большие расходы, ночью
- Строим кластеры, ищем самый большой

А зачем?

- Можем насчитывать признаки
 - Траты «на работе»
 - Траты в «магазине»
 - Траты в «ресторане»
 - Район проживания (соцдем)

Москва. Паркет

- Можно привести к h3_9 и посмотреть, что рядом
- Можно восстановить реальные координаты ;-)
 - Банкоматов
 - Терминалов (вероятностно, но правдоподобно)
- Можно найти центры трат
 - И считать признаки по каждому из центров

Extreme Multi-label

- Probabilistic Label Trees for Extreme Multi-label
- Review of Extreme Multilabel Classification
- Papers With Code
- `napkinxc`
- Развитие идеи факторизации таргета, на основе информационных критериев
- Хорошо работает с разреженными данными

Compressed Sensing

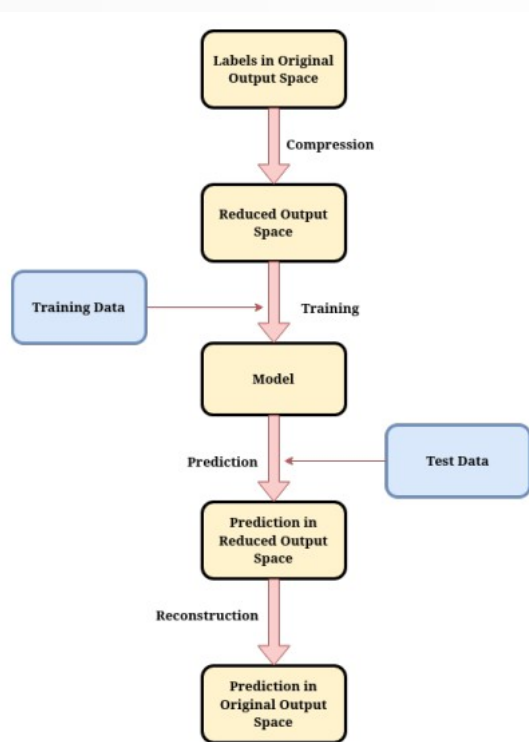


Figure 2: General flow of CS based methods

bit index	representative for labels	bit index	representative for labels
1	{1, 2, 3, 4, 5}	7	{16, 17, 18, 19, 20}
2	{1, 6, 7, 8, 9}	8	{16, 21, 22, 23, 24}
3	{2, 6, 10, 11, 12}	9	{17, 21, 25, 26, 27}
4	{3, 7, 10, 13, 15}	10	{18, 22, 25, 28, 29}
5	{4, 8, 11, 13, 15}	11	{19, 23, 26, 28, 30}
6	{5, 9, 12, 14, 15}	12	{20, 24, 27, 29, 30}

(b)

cluster index	labels in cluster	cluster index	labels in cluster
1	{1, 15}	9	{9, 23}
2	{2, 16}	10	{10, 24}
3	{3, 17}	11	{11, 25}
4	{4, 18}	12	{12, 26}
5	{5, 19}	13	{13, 27}
6	{6, 20}	14	{14, 28}
7	{7, 21}	15	{15, 29}
8	{8, 22}		

<https://arxiv.org/abs/2302.05971>

Deep XML

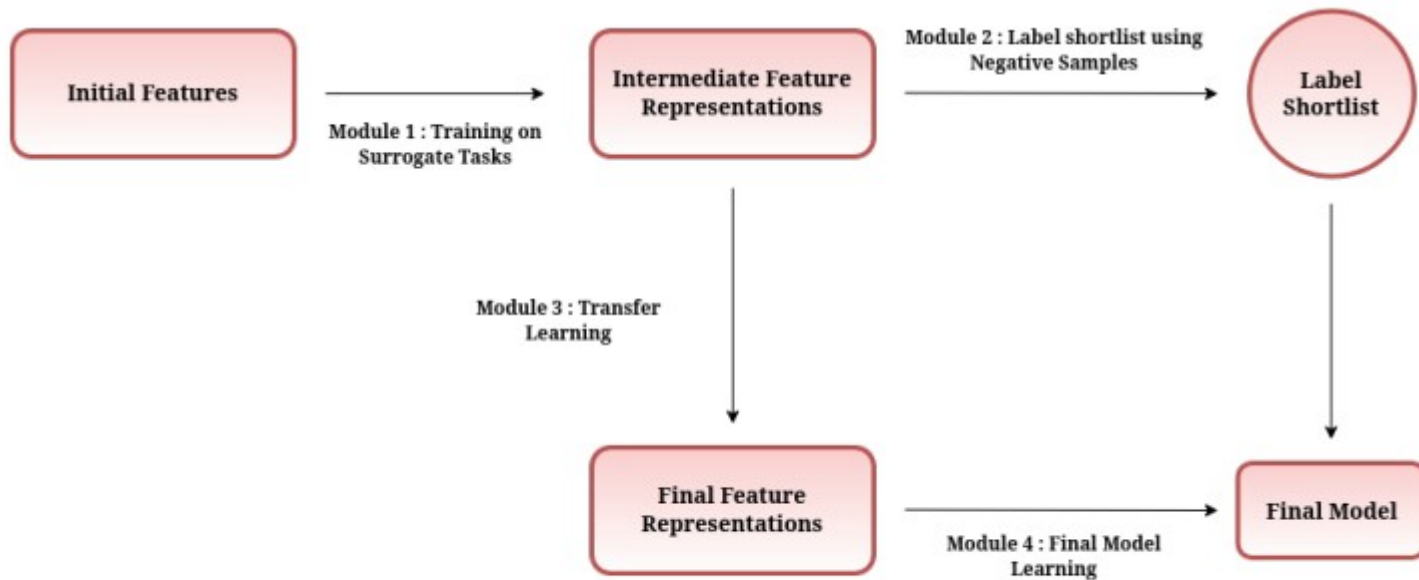
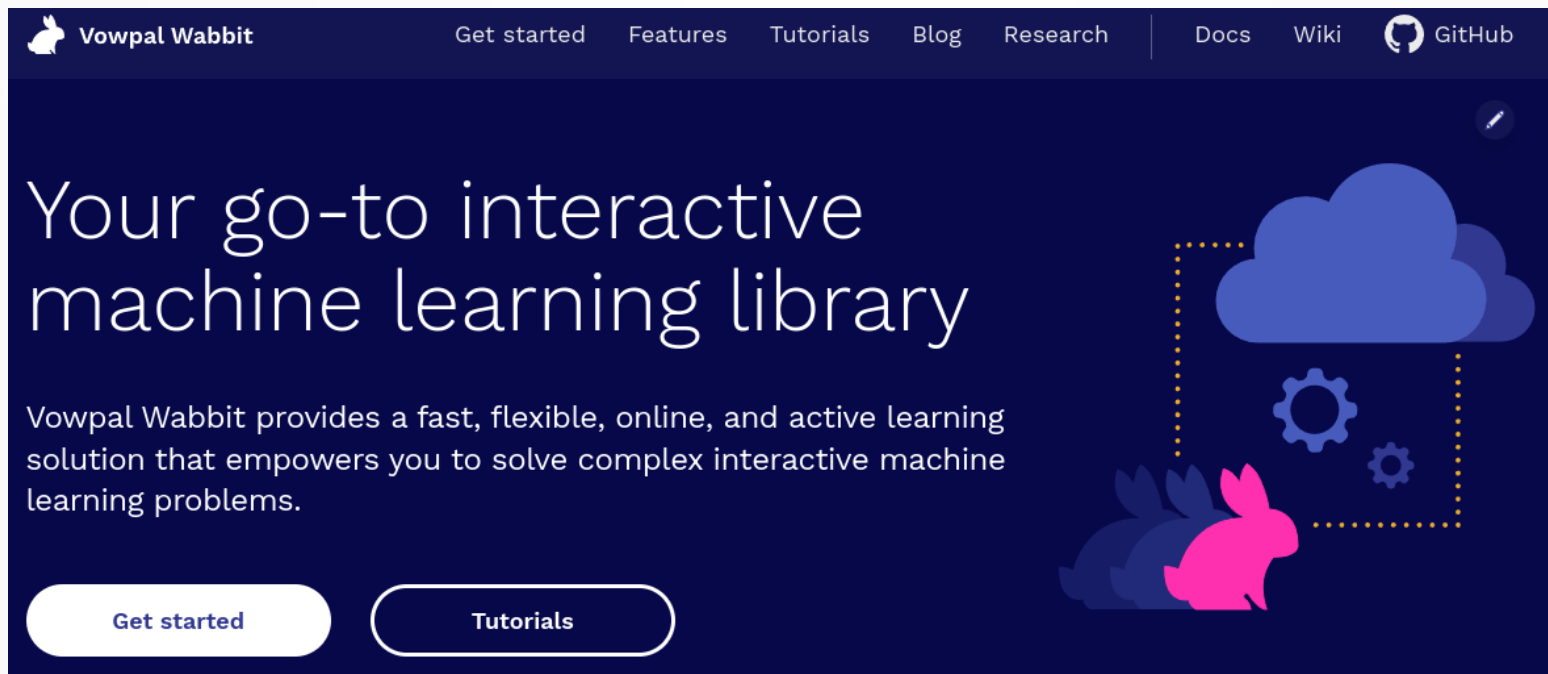


Figure 11: Overview of the DeepXML Framework

BFG 9000

Vowpal Wabbit PLT



The image shows the homepage of the Vowpal Wabbit website. The header features the Vowpal Wabbit logo (a white rabbit silhouette) and the text "Vowpal Wabbit". To the right of the logo are navigation links: "Get started", "Features", "Tutorials", "Blog", "Research", "Docs", "Wiki", and a GitHub icon with the text "GitHub". The main content area has a dark blue background. On the left, the text "Your go-to interactive machine learning library" is written in white. Below this, a paragraph states: "Vowpal Wabbit provides a fast, flexible, online, and active learning solution that empowers you to solve complex interactive machine learning problems." At the bottom left, there are two buttons: "Get started" (white with dark text) and "Tutorials" (white outline with dark text). On the right side of the main content area, there is a graphic illustration featuring a blue cloud, several blue gears, and a pink rabbit silhouette. A dashed yellow line forms a rectangular frame around the cloud and gears.

Vowpal Wabbit Get started Features Tutorials Blog Research Docs Wiki GitHub

Your go-to interactive machine learning library

Vowpal Wabbit provides a fast, flexible, online, and active learning solution that empowers you to solve complex interactive machine learning problems.

[Get started](#) [Tutorials](#)

Размерность признаков

- Разреженные данные, можно учить эмбединги
- С эмбедингами умеет работать бустинг
- На эмбедингах хорошо работает логрег ;-)
- Например **VaE**

Кликстрим

- Можем ли мы свести задачу к кликстриму?
- Упорядочим события. Но — нет временной метки
- Подойдет любой разумный порядок
- Например, приближение к центру мира
 - Например, Охотный ряд
 - Может быть несколько путей в центр
 - Агрегируем предсказания (пуллинг)