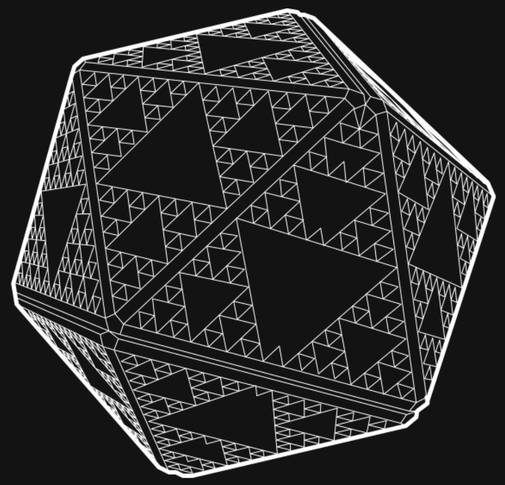
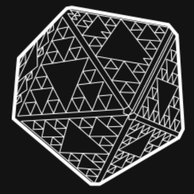


B I G

D A T A



L A B

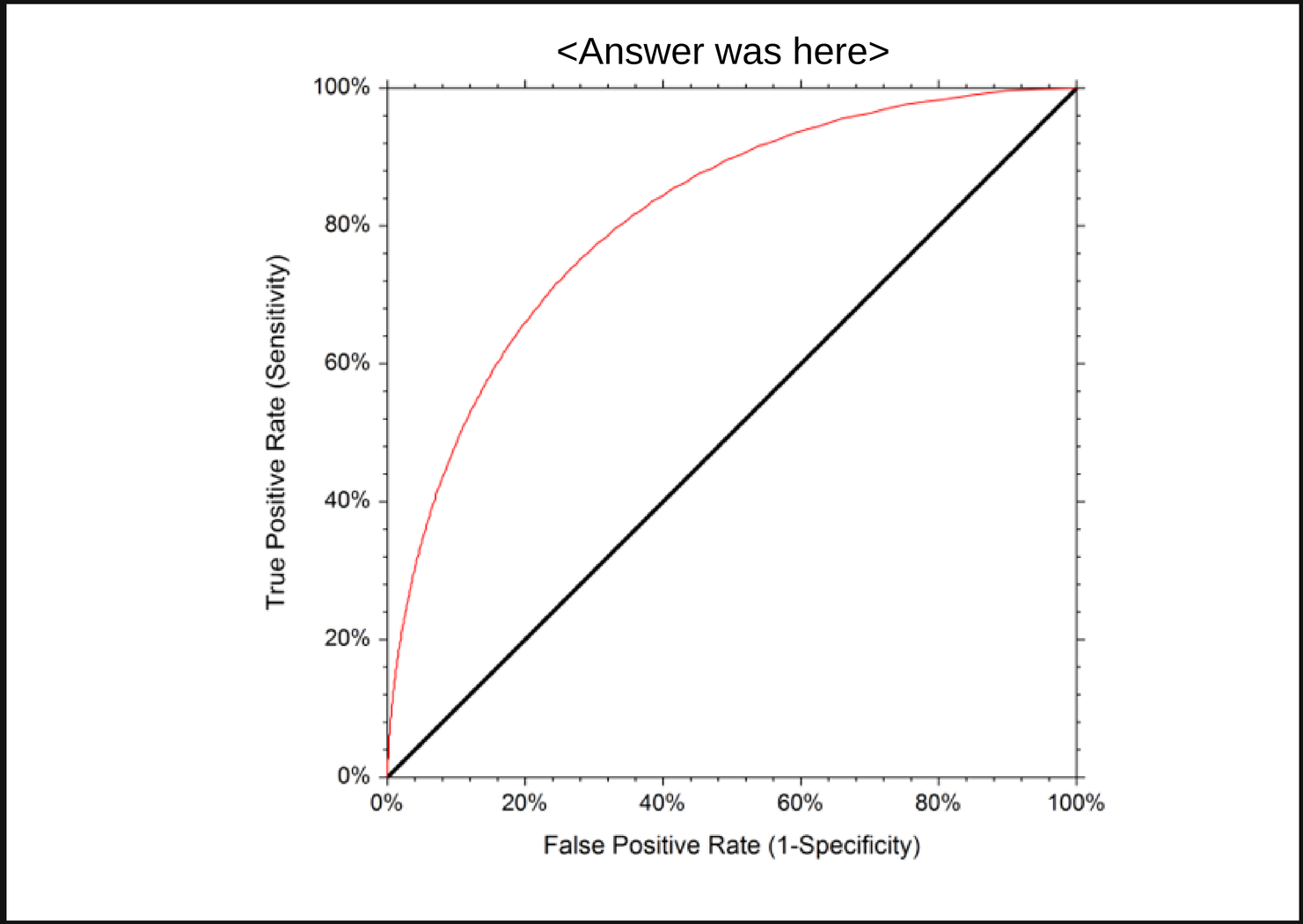
vodafone

Александр Марфин

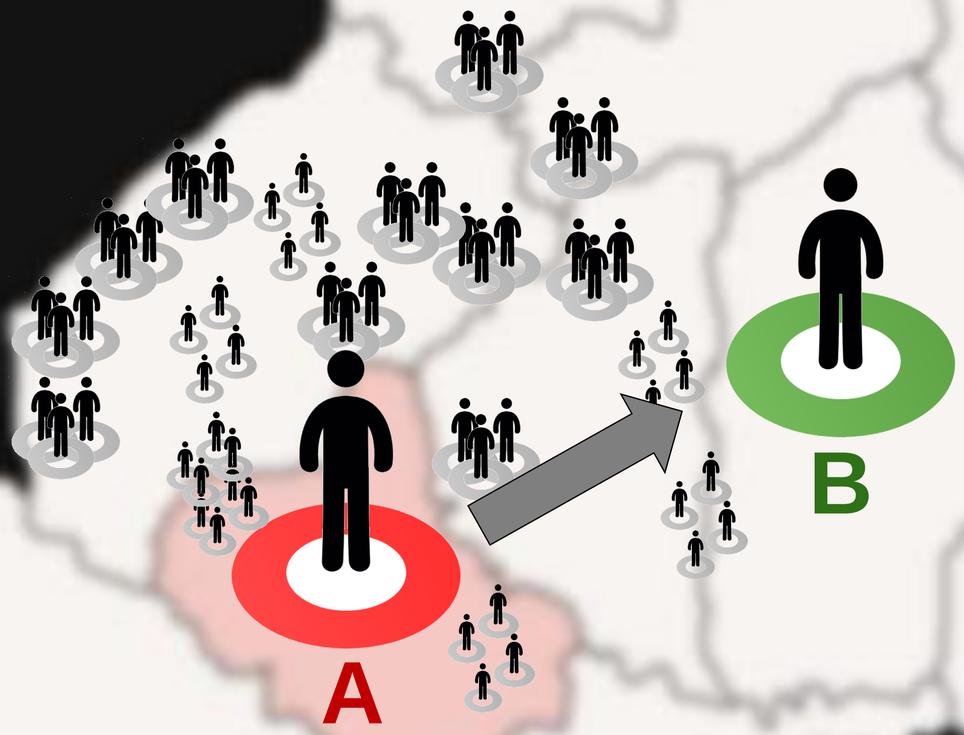
Вступительное слово

руководитель CRM аналитики
Vodafone Украина

Что это такое?



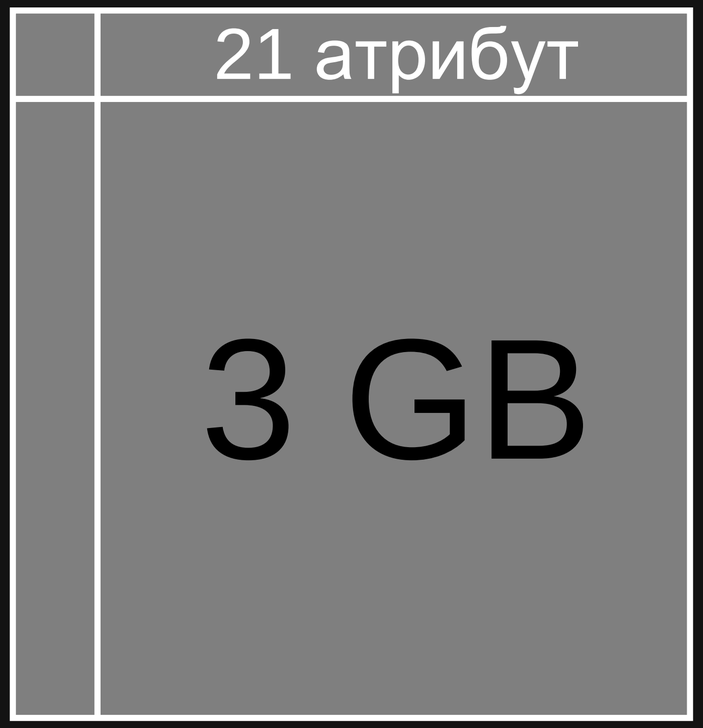
Выборка реальных данных



- Сетевое событие
- Направление
- A номер
- B номер
- Тип события
- Дата и время
- Длительность
- Стоимость
- Координаты A / B
- Тип устройства
- WEB интересы
- Тарифный план
- Категория номера
- Объем интернета
- Сумма пополнения
- Цена устройства

Сколько будет реальных данных?

Начинаем с:



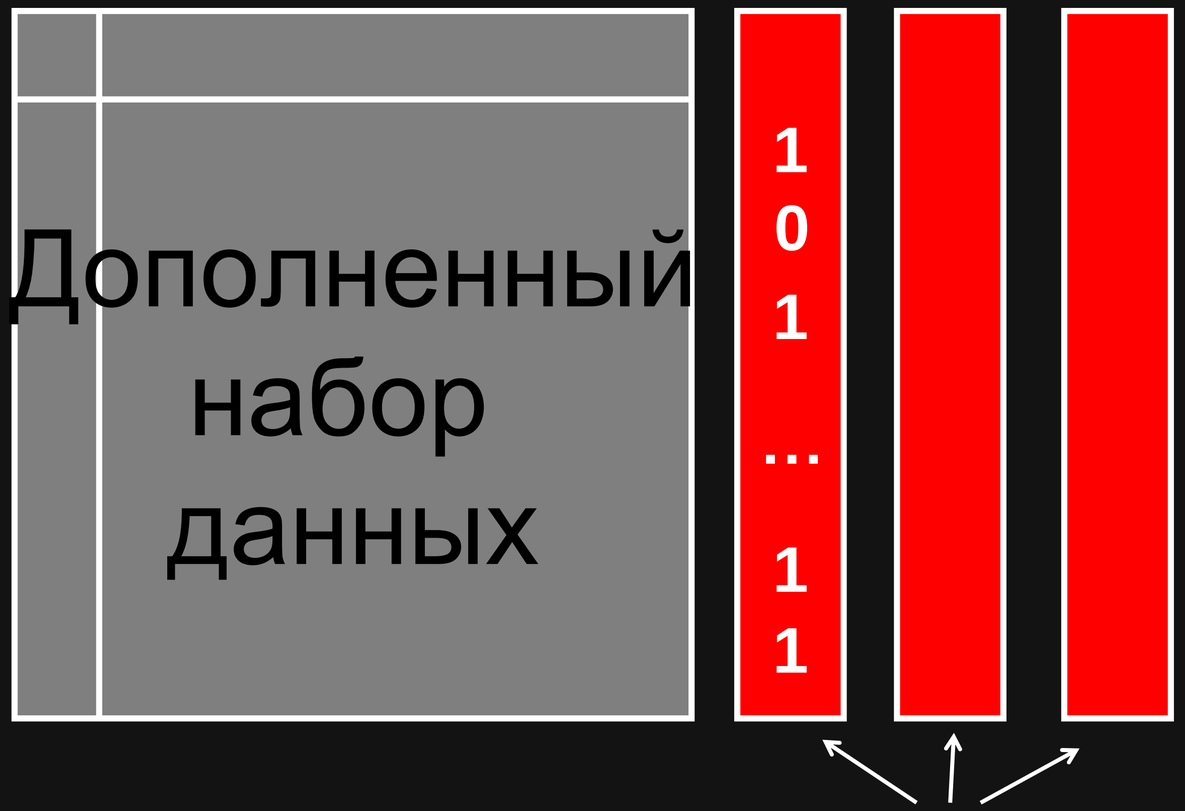
10 000 000

x12

Улучшаем на:

120 000 000

Как проверять гипотезы?



MISC
ROC
SSE
GINI
AUC
Gini index
Logloss
Cumulative Lift

Реальные таргеты от Vodafone Украина и Партнеров

Как отображаются события в выборке?

hash_number_A	event_start_date	LAT	LON	hash_b_number	number_B_category
1877518	18Jul2017 14:17:53	51,29899949	26,53213863	143969	ВУЗ

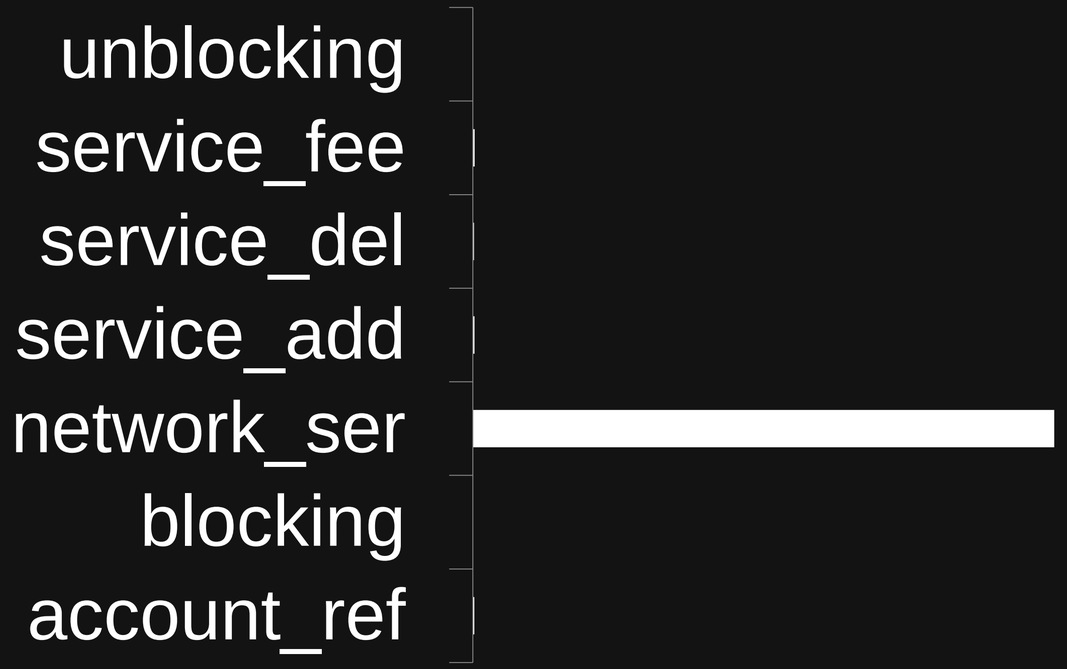
event_start_date	event	event_sub	network_service_direction
18Jul2017 14:17:53	network_ser	pstn	Outgoing

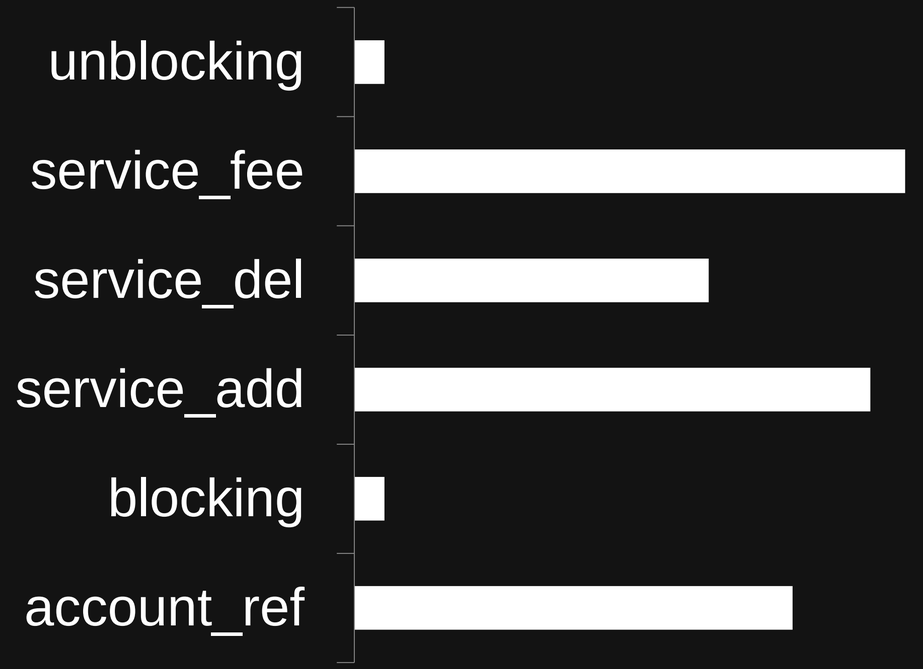
event_start_date	cost	call_duration_minutes	data_volume_mb	hash_accum_code
18Jul2017 14:17:53	1,5000	1	0	0

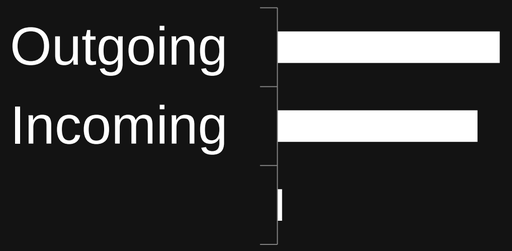
hash_number_A	device_type_rus	phone_price_category
1877518	smartphone	2

hash_number_A	interest_1	interest_2	interest_3	interest_4	interest_5
1877518	Досуг	Наука и образование	Дом и семья	Hi-Tech	Новости и СМИ

Структура данных – event

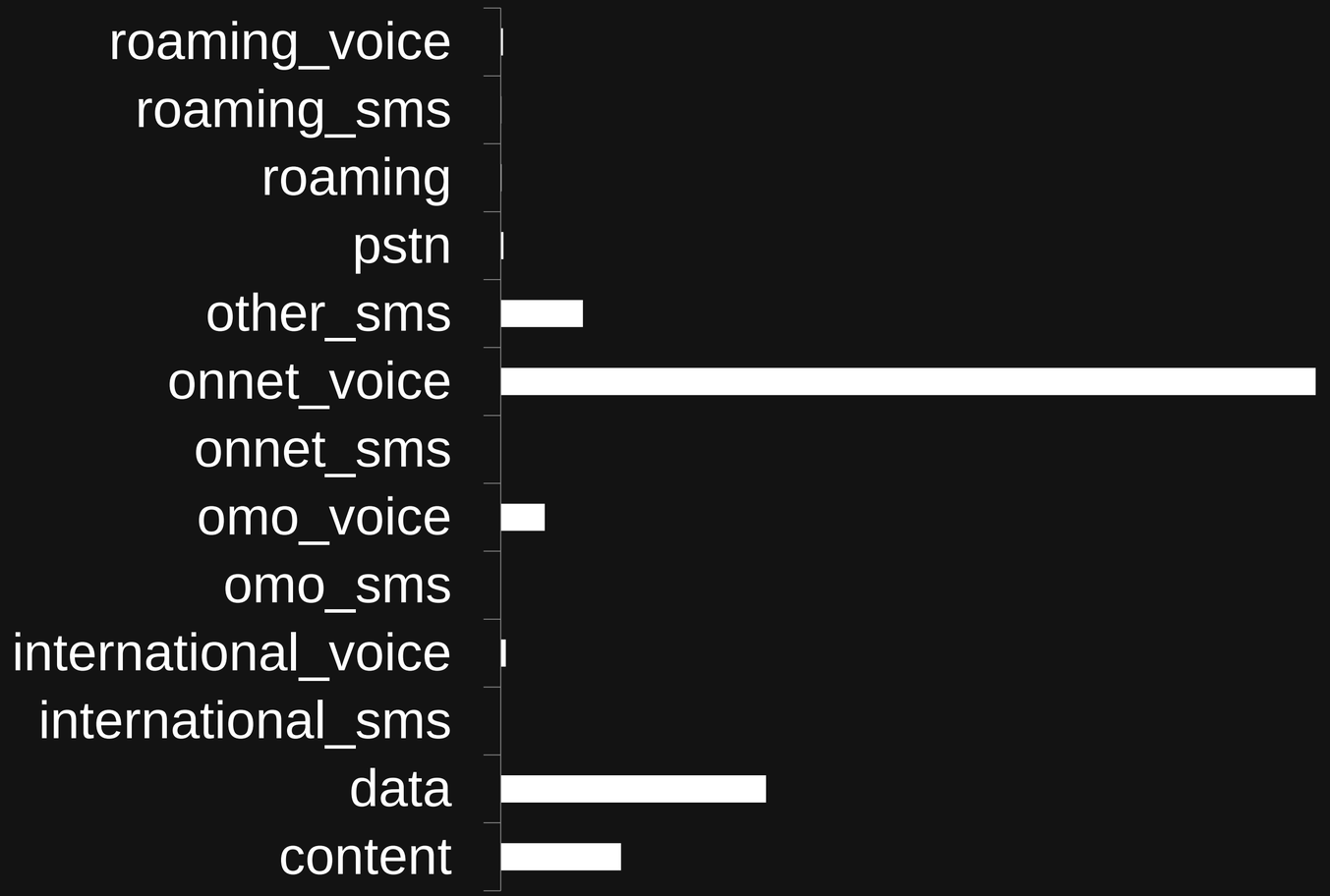




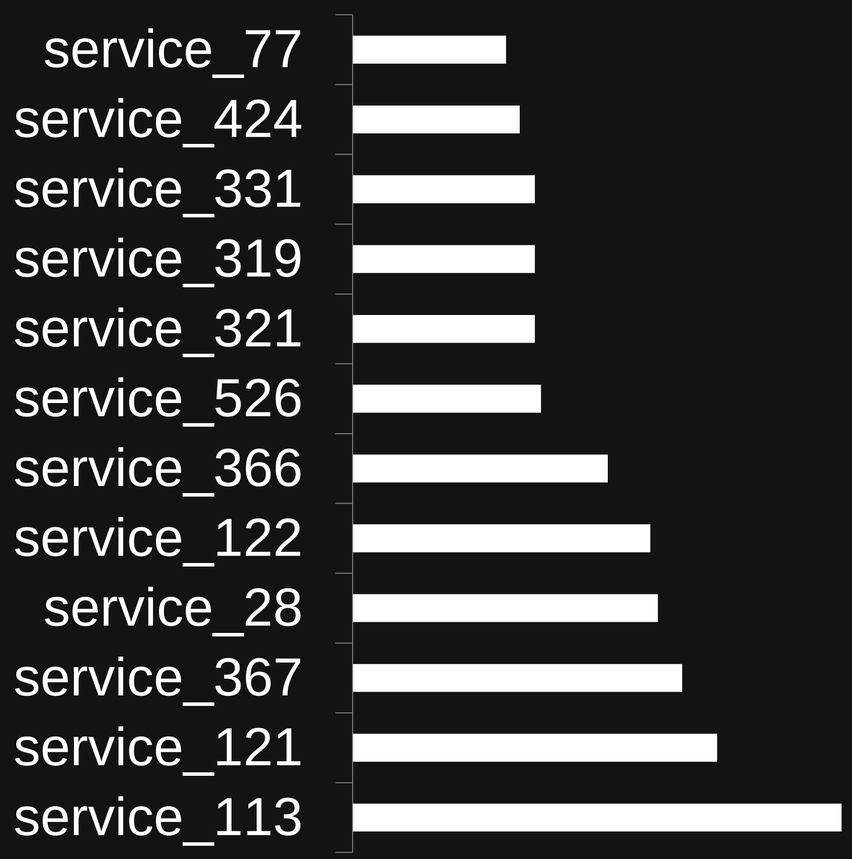


Направление доступно только для сетевых услуг

Event = ‘Network_ser’ -> event_sub



Event = ‘Service_*’ -> event_sub



Предскажем поведение клиента

Будет ли клиент пользоваться мобильным интернетом в будущем?

На вход – Сырая выборка данных
Алгоритм – Logistic Regression
На выход – Binary target
Критерий качества – ROC AUC / ACCURACY

Данные на вход (сырые)

hash_number_A	hash_tariff	event	event_sub	network_service_direction	event_start_date	LAT	LON	cost	hash_b_number	number_B_ca	call_duration_minutes	data_volume_mb
2509423	232437	network_ser	omo_voice	Incoming	07JUN17:15:52...	48.240000254	25.188887999	0.0000	1244324		1.0833333333	0
1874889	231925	network_ser	omo_voice	Incoming	30JUL17:20:32...	50.760638889	25.276360857	0.0000	1244539		7.5333333333	0
1877400	232217	network_ser	omo_voice	Outgoing	24AUG17:16:1...	51.215000763	24.706389652	0.0000	1244669		1	0
1877400	232217	network_ser	omo_voice	Outgoing	11AUG17:18:0...	51.205000763	24.676332698	0.0000	1244669		1	0
1877400	232217	network_ser	omo_voice	Outgoing	13AUG17:09:3...	51.215000763	24.706389652	0.0000	1244669		1	0
1877400	232217	network_ser	omo_voice	Incoming	13AUG17:08:2...	51.215000763	24.706389652	0.0000	1244669		0.2666666667	0
1872815	187131	network_ser	omo_voice	Incoming	18JUL17:12:04...	50.404723748	24.225556064	0.0000	1244751		0.3	0
1872815	187131	network_ser	omo_voice	Incoming	18JUL17:12:04...	50.404723748	24.225556064	0.0000	1244751		0.1	0
1872815	187131	network_ser	omo_voice	Incoming	18JUL17:12:28...	50.404723748	24.225556064	0.0000	1244751		0.1833333333	0
1777541	45807	network_ser	omo_voice	Incoming	23AUG17:14:5...	48.963056318	23.982778286	0.0000	1244840		0.4333333333	0
1777541	45807	network_ser	omo_voice	Incoming	23AUG17:14:4...	48.963056318	23.982778286	0.0000	1244840		1.0666666667	0
1777541	45807	network_ser	omo_voice	Incoming	11AUG17:13:0...	48.941945462	24.745278032	0.0000	1244840		0.2333333333	0
1777541	45807	network_ser	omo_voice	Incoming	10AUG17:16:0...	48.939723239	24.686110221	0.0000	1244840		3.1	0
1777541	45807	network_ser	omo_voice	Incoming	10AUG17:08:4...	48.322498983	25.952778541	0.0000	1244840		0.75	0
1777541	45807	network_ser	omo_voice	Incoming	11AUG17:13:0...	48.941945462	24.745278032	0.0000	1244840		1.0166666667	0
1777541	45807	network_ser	omo_voice	Incoming	11AUG17:03:3...	48.941945462	24.745278032	0.0000	1244840		0.8333333333	0
1777541	45807	network_ser	omo_voice	Incoming	07JUL17:17:25...	48.941945462	24.745278032	0.0000	1244840		1.2333333333	0
1777541	45807	network_ser	omo_voice	Incoming	22AUG17:16:1...	48.638057336	24.944443554	0.0000	1244840		0.3833333333	0
1777541	45807	network_ser	omo_voice	Incoming	22AUG17:16:1...	48.674445716	24.939165777	0.0000	1244840		0.8666666667	0

Target = (data_volume_mb <> 0)

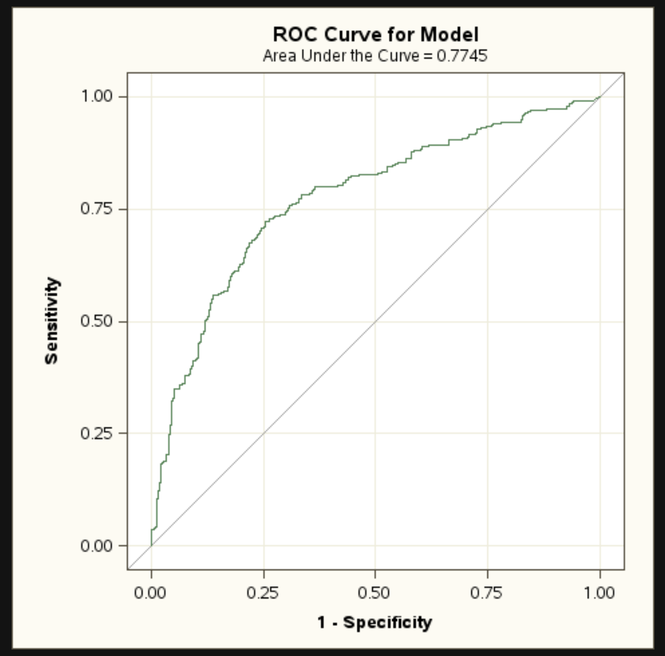
Данные на вход (после агрегации)

hash_number_A	YM201706content	YM201706omo_voice	YM201706onnet_voice	YM201706other_sms	YM201706pst	target
171195	66	21	303	15	1	1
171637	28	0	62	8	0	0
171902	0	0	4	0	0	0
172162	100	7	326	24	0	1
172183	41	0	677	22	0	1
172519	24	1	62	5	1	0
173004	29	1	265	13	0	0
173029	23	0	227	16	0	0
173288	23	1	58	2	0	0
173366	76	5	194	94	1	1
173437	40	10	154	9	0	1
173574	1555	14	382	54	1	1
173781	44	31	116	5	1	0
173953	31	0	77	0	0	0
173986	34	0	4	22	0	0
174055	27	0	507	28	1	1
230394	17	10	631	34	0	1
300179	278	122	443	454	2	1
316821	5	179	374	58	5	1
319685	0	72	674	94	0	1
325089	15	13	156	41	1	1

Target = (data_volume_mb <> 0)

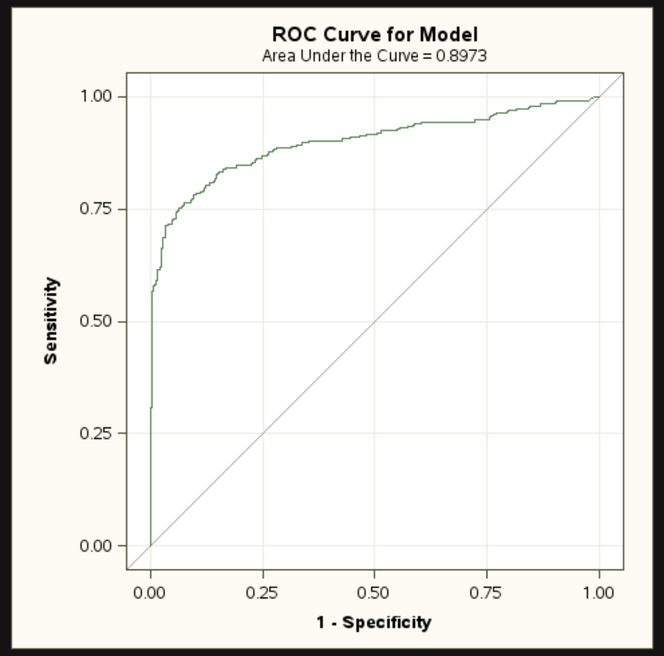
Повышение качества предсказания

Content, Omo_voice, Onnet_voice, Other_sms, Pstn



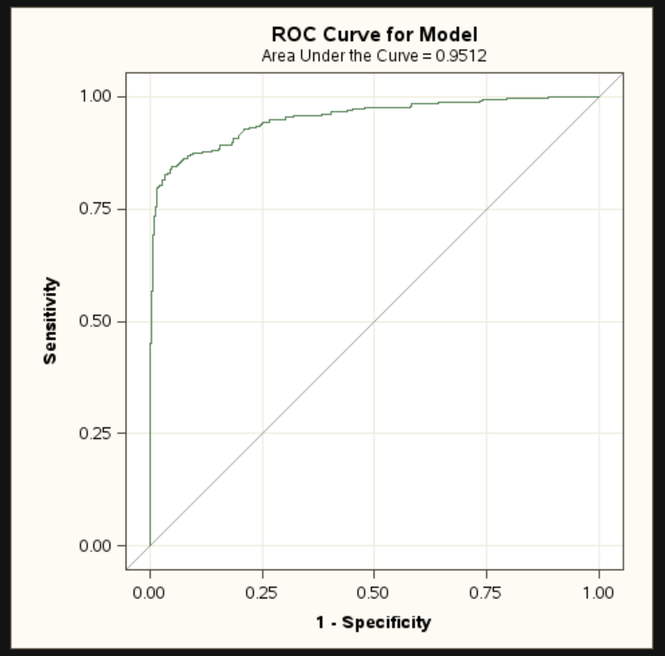
0,774

Content, Data, Omo_voice, Onnet_voice, Other_sms, Pstn



0,897

Content, Data, Omo_voice, Onnet_voice, Other_sms, Pstn Log(1+X)

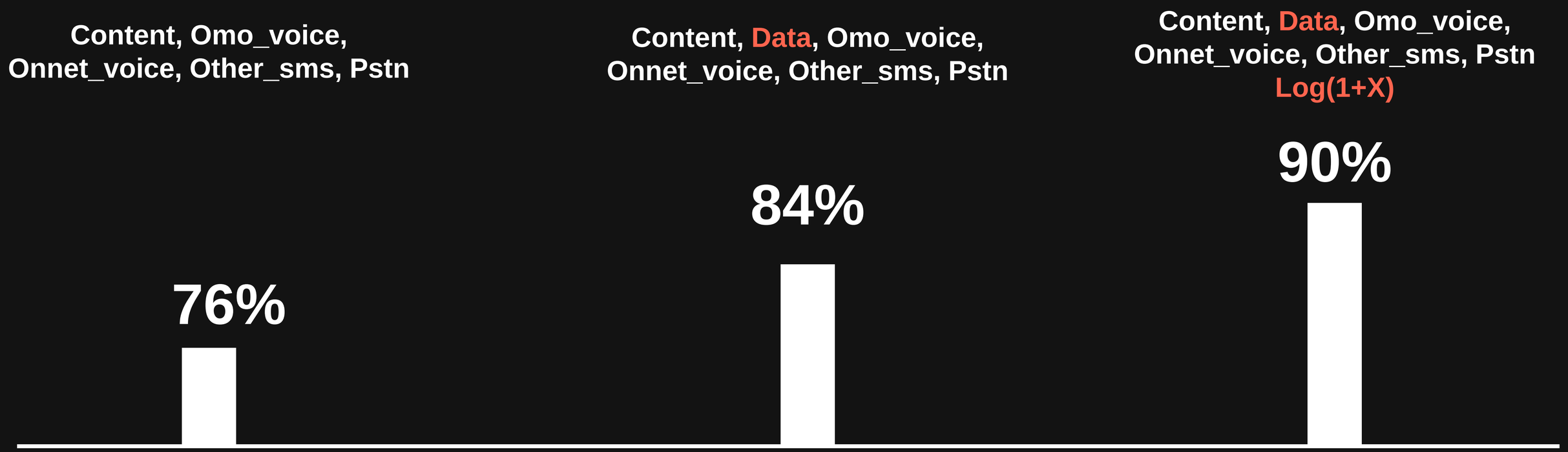


0,951

Проверка стабильности модели

TRAIN 80%

TEST 20%



Можно еще лучше!

- Круто находить не очевидные и не интуитивные правила
- Круто использовать данные разной природы из разных источников
- Круто предсказывать на 3-6-12 периодов будущего
- Круто автоматизировать процесс контроля самообучения моделей и адаптации к новым данным

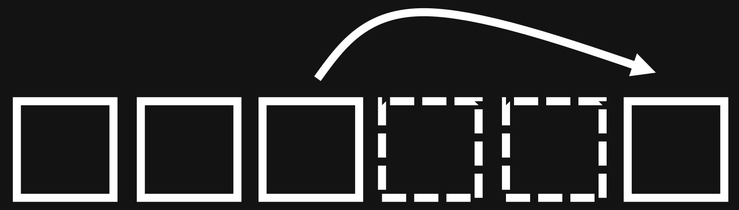
50 Mb

chrome

≠

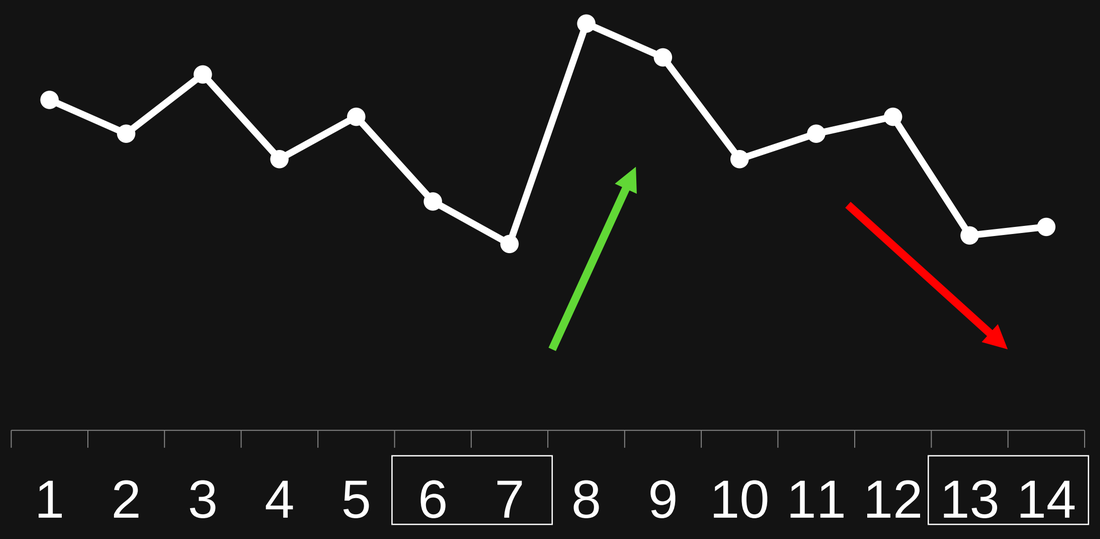
50 Mb

tripadvisor



Творческий поисковый процесс

Используйте бизнесовый и человеческий смысл данных



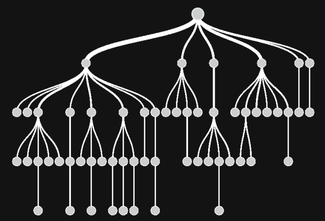
- **Пробуйте разные диапазоны и принципы агрегации данных**
- **Используйте индикаторные переменные**
- **Учитывайте скорость изменения**
- **Ищите аномальные значения**

Инструменты для работы с данными

p2.xlarge instance:
11.75 ECUs,
4 vCPUs, 2.7 GHz, E5-2686v4
61 GiB RAM memory
NVidia Tesla K80(1 GPU - 4992 CUDA)
SSD 50GiB

Rows = 10 000
Columns = 60

Random Forest 15 000 trees



180 секунд

Уникальная Big Data экосистема

- Не надо гоняться по интернету за обучающими выборками
- Здесь можно получить синергию от данных Vodafone Украина и их партнеров
- Проще обогащать данными из открытых источников
- Нет смысла спекулировать на критериях качества (характерно для K*GGLE)
- Результаты труда реально будут улучшать качество жизни населения Украины

Вперед за 99% точностью!

Спасибо! :)