Финальная работа по курсу «Введение в Data Science»

специализация: Data Analyst

Обучающийся: Корнилов И.А.

Skillbox

о себе: Корнилов Иван Анатольевич

Живу и работаю в Екатеринбурге
УРФУ, Прикладная математика (2001г.)
English C2

Работал в ИТ: УГМК, Банк Северная Казна, BSGV, Росбанк, Уралтрансбанк, СВЭЛ, сейчас в УБРиР (с 2020)

На должностях: от Системного администратора до руководителя отдела, сейчас Руководитель Блока Сопровождения ИТ – Систем (в т.ч. Qlickview, HIVE/Hadoop, PowerBI и т.д.)

Постановка задачи «Финальная работа по курсу «Введение в Data Science» специализация: Data Analyst

DA:
Проведите проверку следующих гипотез:
1) Органический трафик не отличается от платного с точки зрения CR (Conversion Rate) в целевые события.
2) Трафик с мобильных устройств не отличается от трафика с десктопных устройств с точки зрения CR (Conversion Rate) в целевые события.
3) Трафик из городов присутствия (Москва и область, Санкт-Петербург) не отличается от трафика из иных регионов с точки зрения CR (Conversion Rate) в целевые события.

Дайте ответы на вопросы продуктовой команды:
1) Из каких источников / кампаний / устройств / локаций к нам идёт самый целевой трафик (и с точки зрения объёма трафика, и с точки зрения CR)?
2) Какие авто пользуются наибольшим спросом? У каких авто самый лучший показатель CR (Conversion Rate) в целевые события?
3) Стоит ли нам увеличивать своё присутствие в соцсетях и давать там больше рекламы?

Постановка задачи «Финальная работа по курсу «Введение в Data Science» специализация: Data Analyst

Анализ сайта «СберАвтоподписка»

В веб-каталоге сервиса на сайте представлены более 20 моделей в наличии. На сайте пользователь совершает целевые действия. Например, нажимает кнопки типа «Оставить заявку», «Заказать звонок». Или нецелевые действия, например, просмотр карточек авто или «блуждания» по основной странице и страницам.

задача — изучить предоставленный датасет, ответить на вопросы и выполнить задание по специализации DA.

Постановка задачи «Финальная работа по курсу «Введение в Data Science» специализация: Data Analyst

Проведите проверку следующих гипотез:
1) Органический трафик не отличается от платного с точки зрения CR (Conversion Rate) в целевые события.
2) Трафик с мобильных устройств не отличается от трафика с десктопных устройств с точки зрения CR (Conversion Rate) в целевые события.
3) Трафик из городов присутствия (Москва и область, Санкт-Петербург) не отличается от трафика из иных регионов с точки зрения CR (Conversion Rate) в целевые события.

Дайте ответы на вопросы продуктовой команды:
1) Из каких источников / кампаний / устройств / локаций к нам идёт самый целевой трафик (и с точки зрения объёма трафика, и с точки зрения CR)?
2) Какие авто пользуются наибольшим спросом? У каких авто самый лучший показатель CR (Conversion Rate) в целевые события?
3) Стоит ли нам увеличивать своё присутствие в соцсетях и давать там больше рекламы?

Выполнение задания:

Проведен разведочный анализ данныех из Google Analytics (last-click attribution model) по сайту «СберАвтоподписка» (файлы GA Sessions (ga_sessions.csv) GA Hits (ga_hits.csv)): проведена чистка, удаление дубликатов,

итоговый результат – данные корректного типа и 0% незаполненных данных

Выполнение задания:

Далее, проведено Объединение дата-сета с событиями (когда кто подключался) и дата-сет с дополнительной информацией о сессиях

Выполнение задания:

Далее, для ответа на вопросы – собраны списки
- все целевые действия и создан новый признак равный целевому действию и не целевому действию
- все типы органического траффика и создаем новый признак равный типу траффика
- все типы устройства и создаем новый признак равный типу устройства
- все города присутствия и создаем новый признак равный городу присутствия

Выполнение задания:

Далее формируем сводные таблицы по признакам из новых типов (целевое / нецелевое), (органический траффик/ не органический)

not_organic	notarget	9808238
	target	59414
organic	notarget	5777695
	target	39872

Выполнение задания:

Далее, используя Биномиальный критерий, расчитываем значимость разницы между долей целевого результата в неорганическом и органическом траффике:

Описание критерия:

Биномиальный критерий

Проверяем значимость доли успешных событий при сравнении двух серий наблюдений
ni — число повторений опыта в i-й серии
mi — число успешных опытов в i-й серии
pi — вероятность успеха в i-й серии
H0
: p1 = p2
HA
: p1 ≠ p2 либо HA
: p1 < p2 либо HA
: p1 > p2
Статистика критерия:

$$T = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2}\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Итоги и выводы:

Вопрос 1) Органический трафик не отличается от платного с точки зрения CR (Conversion Rate) в целевые события.

Нулевая гипотеза : кол-во конверсий среди органического трафика такое же, как и при не-органическом
Альтенативная гипотеза: кол-во конверсий выше при не-органическом траффике Используем биномиальный критерий проверки гипотезы
На уровне значимости 0.05 нулевая гипотеза ПОДТВЕРЖДАЕТСЯ (нет значимой разницы между вариантом А и Б),
т.е. Органический трафик не отличается от платного с точки зрения CR (Conversion Rate) в целевые события.

Выполнение задания:

Аналогично формируем сводные таблицы для типов устройств:

desktop	notarget	3921408
	target	23846
mobile	notarget	11664525
	target	75440

И проверяем с помощью Биномиального критерия значимость доли

Итоги и выводы:

Вопрос 2) Трафик с мобильных устройств не отличается от трафика с десктопных устройств с точки зрения CR (Conversion Rate) в целевые события.

Нулевая гипотеза : кол-во конверсий среди мобильного траффика такое же, как и при траффике с ПК

Альтенативная гипотеза: кол-во конверсий выше при мобильном траффике Используем биномиальный критерий проверки гипотезы
На уровне значимости 0.05 нулевая гипотеза ОТВЕРГАЕТСЯ в пользу альтернативной: Вариаент А более значим (значимая разница Варианта А перед вариантом Б есть)
т.е. трафик с мобильных устройств ОТЛИЧАЕТСЯ от трафика с десктопных устройств с точки зрения CR (Conversion Rate) в целевые события.

Выполнение задания:

Аналогично формируем сводные таблицы для городов присутствия:

city_presence	notarget	9847695
	target	67130
other_city	notarget	5738238
	target	32156

И проверяем с помощью Биномиального критерия значимость доли

Итоги и выводы:

Вопрос 3) Трафик из городов присутствия (Москва и область, Санкт-Петербург) не отличается от трафика из иных регионов с точки зрения CR (Conversion Rate) в целевые события.

Нулевая гипотеза : кол-во конверсий среди трафика из городов присутствия такое же, как из других городов
Альтенативная гипотеза: кол-во конверсий выше из городов присутствия
Используем биномиальный критерий проверки гипотезы
На уровне значимости 0.05 нулевая гипотеза ОТВЕРГАЕТСЯ в пользу альтернативной: Вариаент А более значим (значимая разница Варианта А перед вариантом Б есть)
т.е. трафик из городов присутствия (Москва и область, Санкт-Петербург) ОТЛИЧАЕТСЯ от трафика из иных регионов с точки зрения CR (Conversion Rate) в целевые события.

Выполнение задания:

Для ответа на вопросы продуктовой команды, формируем во-первых сводную таблицу по общему траффику:

event_action_taget_notarget	utm_medium_type_organic_not_organic	device_category_type	geo_city_presence	
notarget	not_organic	desktop	city_presence	1203764
			other_city	349756
		mobile	city_presence	4864114
			other_city	3390604
	organic	desktop	city_presence	1517897
			other_city	849991
		mobile	city_presence	2261920
			other_city	1147887
taget	not_organic	desktop	city_presence	6217
			other_city	1280
		mobile	city_presence	34156
			other_city	17761
	organic	desktop	city_presence	11141
			other_city	5208
		mobile	city_presence	15616
			other_city	7907

Ответы на вопросы продуктовой команды:

1) Из каких источников (кампаний, устройств, локаций) к нам идёт самый целевой трафик (и с точки зрения объёма трафика, и с точки зрения CR)?

С точки зрения объема траффика к нам идет наибольший неорганический траффик с мобильных устройств из городов присуствия
С точки зрения конверсии, к нам идет неорганический траффик с мобильных устройств из городов присутсвия

Выполнение задания:

А также формируем ТОП5 из тех моделей которые наиболее популярны:

taget		
	/audi	168
	/bmw	920
	/haval	565
	/kia	2669
	/lada-vaz	5197
	/land-rover	41
	/lexus	256
	/mercedes-benz	2380
	/mini	70
	/nissan	1239
	/peugeot	131
	/porsche	124
	/renault	1298
	/skoda	7942
	/toyota	1636
	/volkswagen	4867
	/volvo	236

Ответы на вопросы продуктовой команды:

2) Какие авто пользуются наибольшим спросом? У каких авто самый лучший показатель CR (Conversion Rate) в целевые события?

Самые популярные модели с точки зрения наибольшего спроса (ТОП5) это
skoda, lada-vaz, volkswagen, kia, mercedes-benz

Ответы на вопросы продуктовой команды:

3) Стоит ли нам увеличивать своё присутствие в соцсетях и давать там больше рекламы?

В связи с тем, что неорганический траффик НЕ имеет решающее значение для повышения конверсии,
НЕ следует увеличить свое присутвие в социальных сетях, давать больше рекламы.
При этом следует обратить внимание, что наибольшую конерсию мы получаем через мобильные устройства и
из городов присутсвия: следует сконцентрироваться на мобильном сегменте, мобильном приложении и на
городах присутвия, слеует найти самые популярные соц-сети через мобильные устройства и концентрироваться на
тех моделях, что наиболее интересны аудитории (Шкода, Лада-Ваз, Фолькцваген, Киа, Мерседес)
также мобильные интеграции.

Вопросы?