Проект по дисциплине "Научно-исследовательский семинар"
Buzz in social media

Салимова Алина
Киреев Алексей
Экономика и анлиз данных
БЭАД222

2023

Описание датасета

Датасет содержит примеры buzz events ("шумных событий") из социальной сети Twitter.

Данные могут быть получены по ссылке

Датасет представляет из себя 78 колонок, по 7 измерений во времени на показатель, а также последний столбец, по которому будут строиться модели: сколько обсуждений активно через несколько недель после окончания измерений основной части данных

Показатели:

NCD

Количество тредов, созданных за определенный период

AI

Количество новых для темы авторов, взаимодуйствующих с постами

AS(NA)

Нормированный показатель, измеряющий внимание к теме через количество людей

BL

Отношение количества новых тредов ко всем тредам, с которыми взаимодействуют за данный период (новизна темы)

NAC

Общее количество atomic контейнеров созданных по всей соцсети

AS(NAC)

Нормированный показатель, измеряющий внимание к теме через количество взимодействий

CS

"Расползание"взаимодействий - очень скореллировано с BL

AT

Среднее количество авторов, взаимодействующих с постами по теме

NA

Общее количество авторов, взаимодействующих с постами по теме

ADL

Средняя длина дискуссии

NAD

Количество дискуссий по теме

Задачи

1. Предобработать данные
2. Исследовать структуру датасета
3. Построить регрессию, предсказывающую популярность темы спустя время
4. Оценить качество построенной модели с помощью MSE, R^2, выявить лучшую из них.

MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ - это функция риска, соответствующая ожидаемому значению квадрата потерь из-за ошибки

R-квадрат (коэффициент детерминации) - доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными

Процесс работы

(Все описанное может быть наблюдаемо в ноутбуке)

Загрузим датасет по социальной сети Twitter.

Отделим целевую переменную (y, последний столбец) - $MNAD$ от объясняющих признаков (X).

Посмотрим на таблицу попарных корреляций всех переменных X , заметим, что NCD, AI, $AS(NA)$, NAC, $AS(NAC)$, NA, NAD очень сильно скореллированы во времени, чтобы не было проблем с неустойчивостью при построении регрессии из-за скоррелированности регрессоров, каждый из этих параметров примем средним за всё время измерений. Остальные показатели оставляем без изменений - они всё еще содержат разбиение по времени.

Еще раз таблица корреляций, но уже для новых данных. Видим, что все показатели, о которых шла речь выше, скореллированы между собой, усреднять сейчас не стоит, так как показатели имеют разное распредление и смысл. На рис. 1 изображено распределение переменной, которую мы оставим.

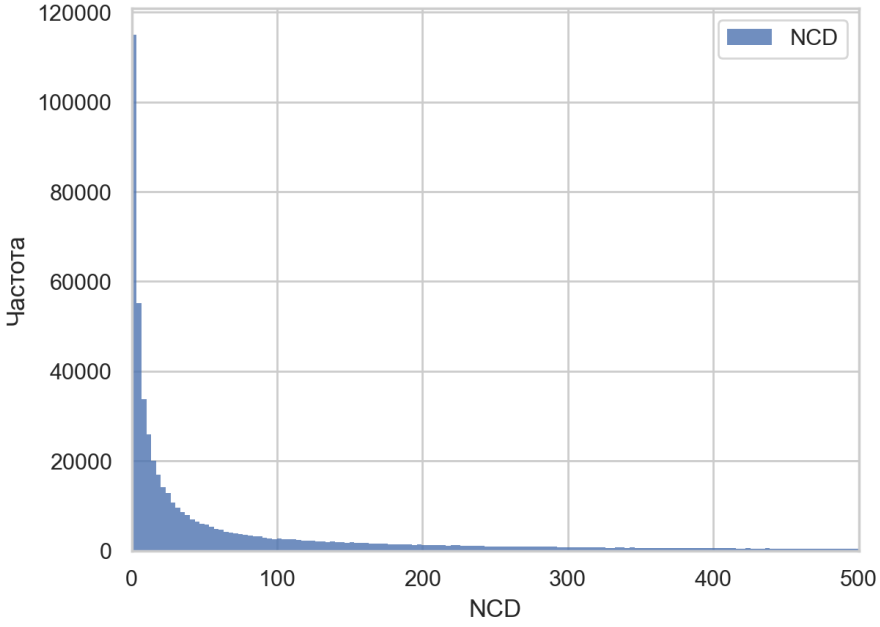

Рис. 1: Гистограмма значений NCD

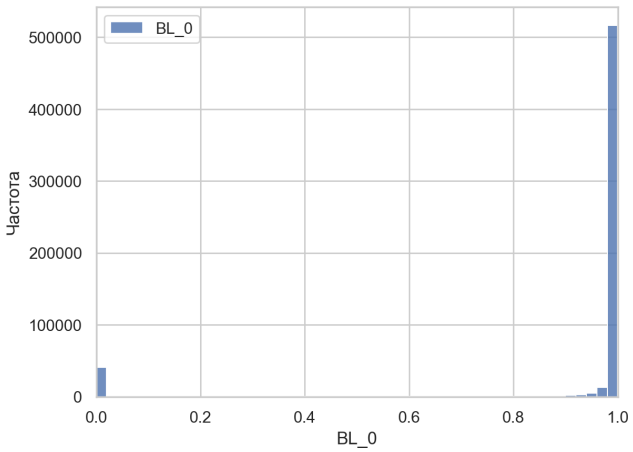

Рис. 2: Гистограмма значений BL

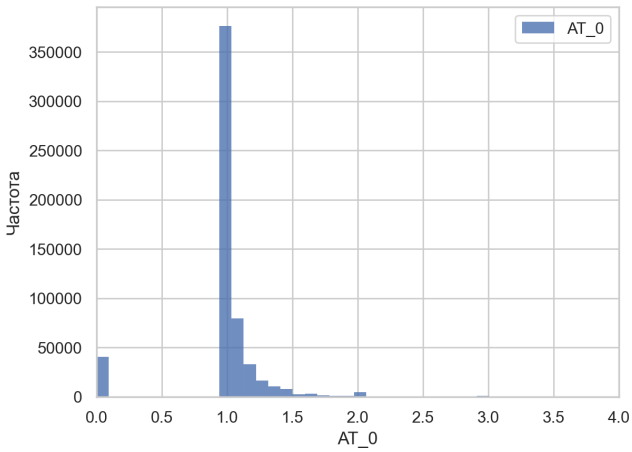

Рис. 3: Гистограмма значений AT

Опять посмотрим на таблицу корреляций, видим что ADT_i и AT_i, а так же BL_i и CS_i очень сильно скорреллированы, так что оставим по одному показателю из каждой пары: BL и AT.

Теперь у нас остались показатели BL_t, AT_t - распределены во времени и один показатель NCD. Распределения BL и AT показаны на рис. 2 и 3, в целом для всех периодов они похожи

Во-первых, отметим, что данные распределены неравномерно, большая их часть имеет очень похожие (почти одинакоые) значения по всем параметрам.

Во-вторых, начнем строить регресию. Разделим данные на тренировочную и тестовую выборки, тренировать модель будем на первой, оценивать ее качество - на второй. Самая наиваная регрессия, куда мы добавим вообще все переменные, что у нас остались. Получаем $MSE = 56917.27$

Каждый раз будем отбрасывать по одному регрессору с $p\ value$ меньше 0.05, всё идет нормально, $R^2 \approx 0.9$, но как только мы выкидываем NCD, оно резко падает до 0.1, тогда выкидывать его мы не будем. Кстати, интересно, что мы в итоге убрали 6 переменных, но R^2, округленное до сотых не изменилось. Теперь $MSE = 56903.12$, упало, но очень незначительно. Убрали регрессоры, состояние не ухудшилось, супер, пока оставляем так.

Из-за несбалансированности данных некоторые тесты линейной регрессии ($p\ value$), которым мы пользовались раньше, могут немного нас обманывать, как произошло с NCD. Опираясь на то как сильно падает предсказательная сила модели, при убирании NCD, поппробуем построить модель только с ней, получаем

$$MNAD = 1.09 \cdot NCD$$

На ней $MSE = 56888$ - еще ниже чем была раньше, супер. Оставляем такую модель. Она интерпритируется очень легко: чем больше популярность треда сейчас, тем больше она будет в будущем, посмотрим на то как соотносяться предсказанные данные и реальные.

Видим, что есть совсем мало точек, которые нами плохо предсказываются, причем они появляются только при достаточно высоких значениях $MNAD$ - скорее всего, это темы, резко поймавшие хайп, что предсказать невозможно. Основную часть тем мы предсказываем достаточно хорошо. Так же отметим, что количество выборосов - единцы, хотя в датасете содержаться сотни тысяч строк.

В заключение, наша оценка достаточно точная, модель получается не слишком сложной. Во многом это согласуется с тем, что пишут авторы, они говорят о том, что можно сделать неплохую модель, но она обладает некоторыми несовершенствами из-за несбалансированности данных, это и удалось пронаблюдать на гистограммах.