ВШЭ БЭАД222

Блохин Павел, Бахишев Никита

Анализ популярности новостей в разных социальных сетях

11 ноября 2023 г.

Содержание

1 Вступление

1.1 Описание

Во время нашего проекта мы исследовали датасеты, содержащие информацию об оценки популярности новостей в трех социальных сетях: Facebook, GooglePlus и LinkedIn. Мы попытались предсказать "level of popularity"четырёх тем для новостей (Obama, Palestine, Economy, Microsoft) на основе предыдущих данных, используя модели для анализа временных рядов. Всего мы разработали 2 модели линейной регрессии.

1.2 Цели и методы

Цели:

- Проанализирвав датасет, выбрать наиболее активную социальную сеть по охвату новстей
- Обозначить самую популярную и менее популярную тему в соцсетях
- Построить модель, предсказывающую итоговый интерес пользователей к новости на основе заданных временных промежутках

Методы:

- библиотеки для визуализации:
    1. seaborn
    2. matplotlib
- библиотеки для подсчётов и анализа:
    1. pandas
    2. sklearn
    3. statsmodels

2 Предобратка данных

В самом начале работы с нашеми датасетами, мы решили разбить их на группы по темам новосетй. После мы заметили что в итоговом датасете (news_final), содеражщем все темы и итоговые оценки, было больше новостей чем в отдельных датасетах, которые содержали изменение популярности каждой темы в разынх соцсетях. Это означало лишь то, что эти новости не были опубликованны в наших соцсетях, поэтому мы их изключили.
После перевели строковые значения столбца "даты публикации"в формат pd.date, чтобы было возможным группировать и производить вычесления.
На этом обработка датасето закончена, можно преступать к первичному анализу данных.

3 EDA

3.1 Соцсети

Для начала мы решили посмотреть в каких социальных сетях новости придаются большей огласки. Для этого мы группировали подготовленные датасеты(разбитые по темам) по дате и смотрели на средную итоговую оценку в каждой соцсети, построили графики для наглядности. Из полученных данных мы выяснили, что в Facebook новости набирают самый большой рейтинг, а в GooglePlus самый низкий. Также по этим графикам можно увидеть на сколько популярны самы темы, но смотреть на это не удобно, поэтому разберем это в следующей части

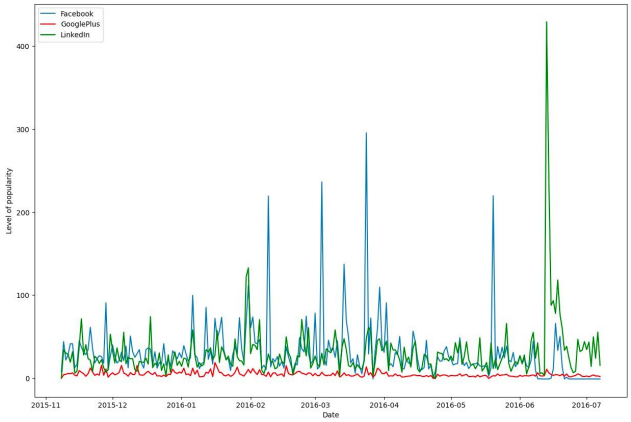

Рис. 1: Microsoft

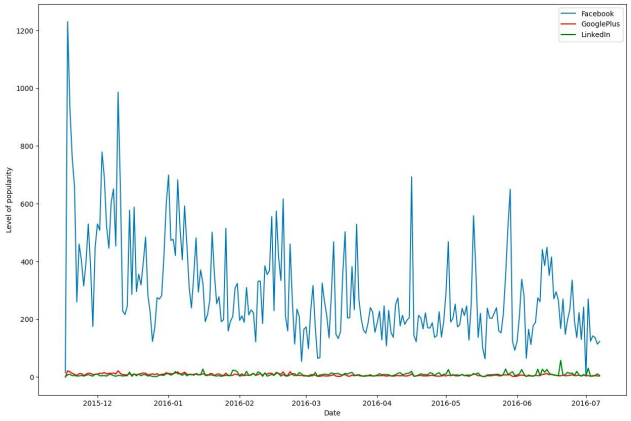

Рис. 2: Obama

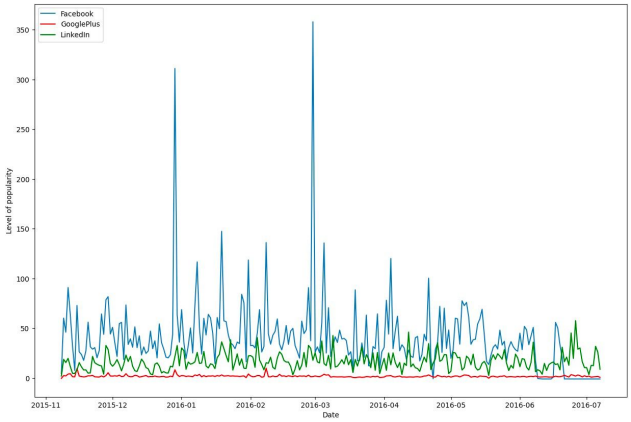

Рис. 3: Economy

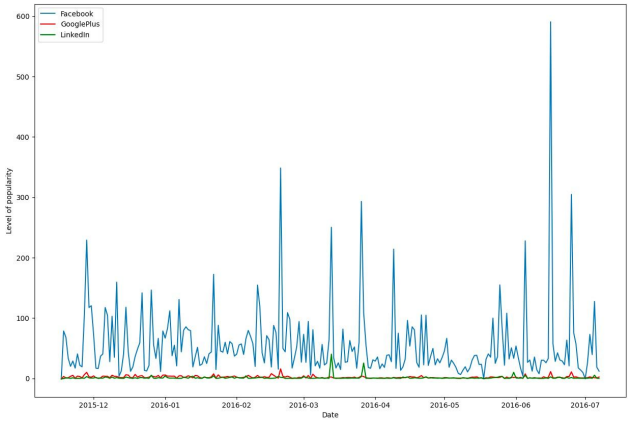

Рис. 4: Palestine

3.2 Темы новостей

Для того чтобы посмотреть на то, какая тема более популярная мы объединили обработанные датасеты, группировали их по темам и смотрели средний балл в каждой соцсети. Далее мы вывели тепловую карту чтобы наглядно увидеть какой рейтинг набирают темы на всём промежутке времени.

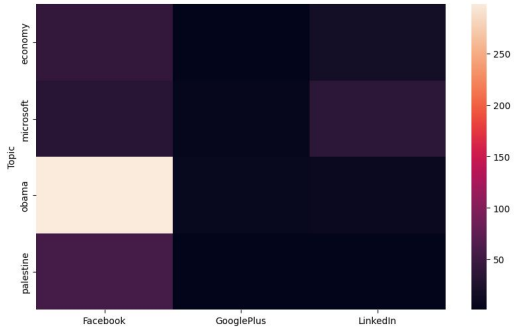

Рис. 5: Тепловая карта 1

На первой карте сразу бросается в глаза то, что тема "Obama"в Facebook имеет очень большой рейтинг (запомним), но из-за этого очень плохо видна разница между другими датасетами, надо это как-то исправть.

Так-то лучше. Поиграв с ограничениями, нашли то самое(60) где также выделяется тема "Obama Facebook"как самая лучшая, но и отличия между другими темами стало заметным. Мы видим, что тема "Palestine"является самой непопулярной, что мы показали в вычислениях, а "Obama"является хитом.

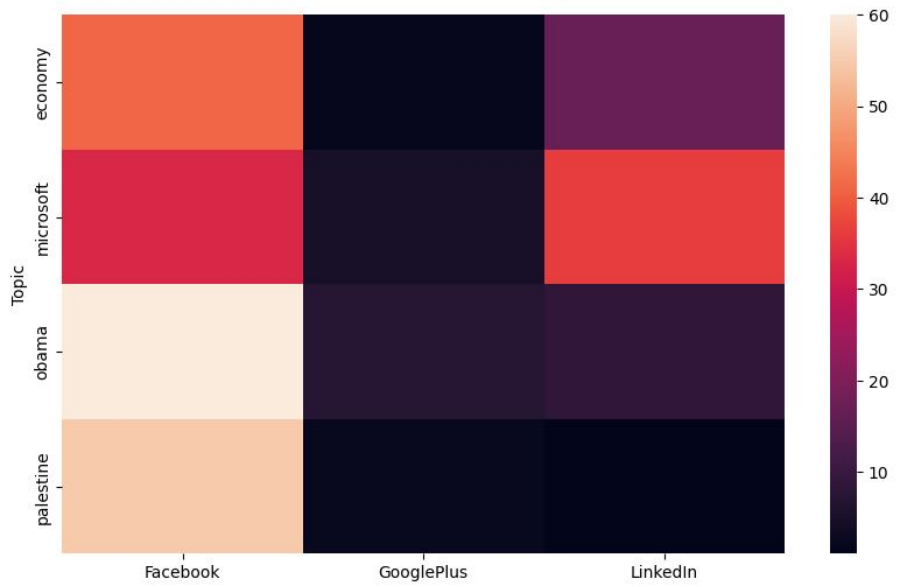

Рис. 6: Тепловая карта 2

Также по этой карте хорошо видны в каких социальных сетях какая тема актуальна.

4 Модели

4.1 Наивная линейная регрессия

Для начала мы решили построить простой вариант линейной регрессии, которая бы брала в качестве признаков каждый пятидесятый столбец датасета с временным рядом и предсказывала итоговый интерес пользователя спустя двое суток после публикации. Брать каждый столбец, отвечающий за временной промежуток в 20 минут было бы не очень хорошо, так как можно заметить, что значения интереса обычно остаются статичным на период, за который отвечают несколько столбцов, то есть TS144 (последние 20 минут и итоговый интерес) почти не отличается от непосредственно предшедствующих ему. Мы простроили ее на примере facebook, попытавшись обучить 4 модели для соответствующих тем (обама, палестина, экономика и майкрософт).
Результаты получились достаточно высокими, модель с очень хорошей точностью смогла проявить себя на тестовой выборке, значение р квадрат было очень близко к единице (все параметры линейной регрессии выведены в ноутбуке).

4.2 Не наивная регрессия

Брать каждый 50-й столбец достаточно опрометчиво, поскольку в таком случае мы ничего не знаем про линейную зависимость между значениями, поэтому мы решили пойти по более серьезному пути.
Была построена матрица корреляций и так же были обучены две модели линейной регрессии, одна из которых использовала ограничение на корреляцию между признаками (нашими столбцами) как 0,9, а другая как 0,95. Это означает, что вы берем столбцы, которые коррелируют между собой на меньше, чем вышесказанные значения. Для этого были отобраны только те столбцы, которые удовлетворяли бы заданным ограничениям. Так мы попытались избежать нестабильных решений, отобрав нужные нам столбцы, опираясь на матрицу корреляций.
Получились две модели, первая из которых (с ограничением в 0,9), не очень хорошо справилась в предсказанием итоговой переменной: р квадрат 0.4823, среднеквадратичная ошибка порядка 44 тысяч против полутора при наивной реализации. Вторая же модель с ограничем в 0.95 лучше справилась с поставленной

задачей: p квадрат 0.89, среднеквадратичная ошибка порядка 9 тысяч, что говорит нам о высокой эффективности предсказаний модели, не потеряв при этом стабильность решений. На рисунке ниже, можно посмотреть, как визуально модель справилась с задачей:

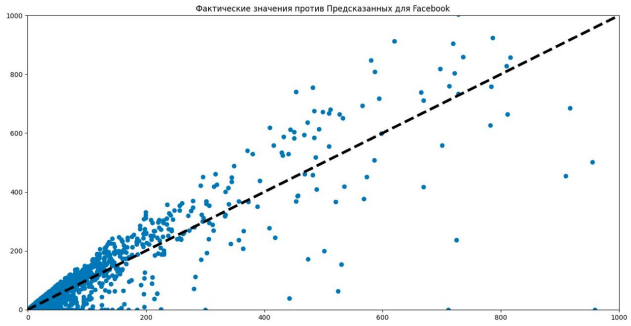

Рис. 7: Фактические результаты против предсказанных

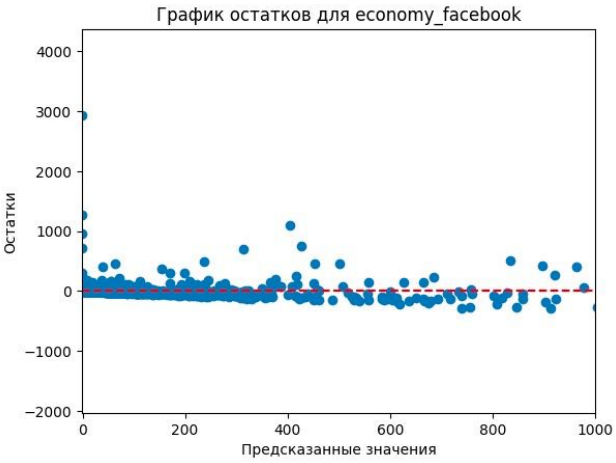

Рис. 8: График остатков

На первом графике синяя точка имеет координаты X, Y, где Y - предсказанное значение, а X - фактическое. В идеальной модели все точки соответствовали бы графику $y = x$, модель хоть имеет разброс, но достаточно часто находится в окресности "идеальной прямой".
Второй график показывает остатки и можно заметить, что значения имеют достаточно небольшой выброс.

Таким образом, была успешно построена и протестирована модель с более рациональным выбором признаков.

4.3 Итоги

Мы выполнили поставелнные задачи, обработали датасеты, провели разведочный анализ и построили 2 варианата линейных регрессий, и выявили какая из них является более эффективной, несмотря на сложные числовые ряды.