Рассмотрение признаков, выводы из визуального анализа:

float data

1. page_likes_num (done: убрано, чтобы ускорить алгоритм, этот признак почти не влияет на результат)
- можно попробовать прологарифмировать (стало хуже на тысячные)
- можно попробовать убрать выбросы лучше после логарифмирования (стало хуже на тысячные)
- корреляция с целевой переменной слаба (0.06 по Пирсону), можно попробовать убрать (стало лучше на тысячные)

2. page_checkins (done: убрано, чтобы ускорить алгоритм, этот признак почти не влияет на результат)
- слишком много нулей (больше половины: 25 тысяч из 40)
- может быть стоит убрать из рассмотрения из-за нулей
- может стоит перекодировать (0 против всех)
- корреляция с целевой переменной слаба

3. page_talking_about (done: логарифм от +0,05)
- может стоит прологарифмировать, но при этом стоит обработать -inf, например np.log(data['page_talking_about']+0.05) (стало лучше на сотые)
- возможно, стоит попробовать убрать этот признак, корреляция с целевой переменной слаба

4. page_statistics' 5-29 cols (done: убрано 10, 15, 16, 17, 18, 19, 21, 23)
- 'page_statistics_10' и 'page_statistics_23' очень слабо коррелируют с целевой переменной (их можно убрать)
- между некоторыми значениями наблюдаются сильные корреляции (например 0 и 15 - по модулю больше 0,99), скорее всего стоит убрать: 10, 15, 16, 17, 18, 19, 21, 23 (более 0,99) (стало на несколько тысячных хуже) и возможно стоит убрать: 3, 7, 12, 14, 24 (более 0,95) (оставить)
- признаки 0, 5, 10 и 15 имеют слишком много нулей (больше 37 тысяч из 40). Их лучше убрать, либо один против всех (10 и 15 убраны, если убрать 0 и 5 - оценки уменьшаются на тысячные, если перекодировать - средняя оценка такая же, отклонение стало меньше, лучше будет убрать, так как дисбаланс очень большой)
- признак 23 имеет половину 0, можно сделать один против всех (признак убран)
- для 1,2,11 можно попробовать корень или логарифм распределения (оценки стали немного лучше)

5. 'comments_' (done: убран признак comments_num_in_first_24_hours, один против всех comments_num_in_last_48_to_24_hours)
- 'comments_num_before_base_time' сильно коррелирует с 'comments_num_in_first_24_hours', можно попробовать без одного из этих полей (убрать: comments_num_in_first_24_hours)
- во всех наблюдается большое количество нулей, может стоит кодировать (один против всех, там где 15-20 тысяч и другое кодирование для других, но из-за больших корреляций с целевой переменной может быть не очень удачно)

- можно попробовать логарифм (не сильно повлияло на распределение)

6. character_num_in_post (done: убрать)
- Целочисленное значение, которое можно попробовать закодировать (до 100 знаков и тд)
- Много нулей: чуть меньше 5 тысяч из 40
- Наблюдается закономерность с целевой переменной: чем больше символов, тем меньше лайков чаще всего (а корреляция очень слаба)

7. share_num (done: оставить как есть)
- Нулей нет, много единиц: более 8 тысяч из 40, можно попробовать перекодировать по диапазонам (нет сильного влияния на оценку)
- Логарифм не сильно помогает из-за большого числа единиц

cat data

8. 'page_cat' (done: убрать)
- 81 категория
- каких-то значений много 7491, каких-то мало 1, возможно, признак не стоит рассматривать, либо стоит закодировать по частоте (при пороге 1500 оценки стали, при 2000 оценки стали незначительно хуже на тысячные)
- в pdf файле есть текстовые названия, названия содержат однокоренные слова (sport, art), можно попробовать рассмотреть как тексты (стало хуже на тысячные)
- можно закодировать one-hot (улучшилось, но не намного)
- можно и тексты, и one-hot (стало хуже на тысячные)

9. base_time (done: one-hot)
- В документации указано, что это закодированное значение в диапозоне от 0 до 71, на практике от 0 до 72
- Разница в объеме данных для каждого значения небольшая (минимум 489 записей, максимум 625 записей)
- Распределение целевой переменной для некоторых значений могут сильно различаться
- Можно попробовать развернуть в one-hot encoding (оценки значительно улучшились), но тогда возможно будет потеряна связь между значениями признака, а зависимость целевой переменной от признака к признаку заметно: чем больше признак, тем меньше разброс значений переменной.

10. h_local (done: one-hot)
- значения от 1 до 24, очень плохо сбалансировано, почти все значения равны 24. Можно попробовать убрать или сделать один против всех или one-hot (стало немного лучше)
- зависимость целевой переменной между соседними признаками не наблюдаема, корреляция с целевой переменной мала, но это категориальный признак

11. post_published_weekday_ (done: убрано)
- для каждого столбца 5-6 тысяч ненулевых значений

- для всех столбцов boxplot целевой переменной для значений 0 и 1 выглядит одинаково,но очень большое количество данных отсеиваются как выбросы, поэтому стоит посмотреть на матрицу корреляций
- из матрицы корреляций видно, что признаки очень слабо коррелируют с целевой переменной, можно попробовать без них

12. base_datetime_weekday_ (done: убрано)
- для каждого столбца 5-6 тысяч ненулевых значений
- для всех столбцов boxplot, кроме пятницы и субботы целевой переменной для значений 0 и 1 выглядит одинаково,но очень большое количество данных отсеиваются как выбросы, поэтому стоит посмотреть на матрицу корреляций
- из матрицы корреляций видно, что признаки очень слабо коррелируют с целевой переменной, можно попробовать без них

target

- очень много нулей: 22 тысячи из 40

Возможные дополнительные признаки:

1. Из признаков post_published_weekday_ и base_datetime_weekday_ вывести бинарный признак выходного дня
2. Отношение share_num к comments_num_before_base_time
3. Отношение page_talking_about к page_likes_num #доля репостов к лайкам
4. Отношение comments_num_in_last_24_hours к comments_num_before_base_time
5. Отношение comments_num_in_last_24_hours к comments_num_in_last_48_to_24_hours #скорость падения интереса
6. Отношение 'comments_num_in_last_24_hours' к 'comments_num_in_first_24_hours'
7. Попробовать квадратичные признаки

Выводы для новых признаков:

1. **is_post_published_in_holiday и is_base_datetime_holiday** - очень много нулей: 30 тысяч из 40. Из scatter plot можно увидеть небольшую зависимость целевой переменной, стоит посмотреть на корреляцию. Корреляции слабые. (результаты ухудшились на тысячные) (done: не добавлены)
2. **share_part_in_comm** - можно попробовать взять логарифм, корреляция слаба (результаты ухудшились на тысячные) (done: не добавлен)
3. **page_share_part_in_likes** - стоит попробовать рассмотреть корень 4 степени. С этим признаком скорее всего можно работать.(реультаты худшились на тысячные) (done: не добавлен)
4. **comm_part_1** - 15 тысяч нулей из 40, получилось мультимодальное распределение -> стоит разделить на 2 части бинарно: является 0 или нет (done: добавлен comm_part_1)
5. **comm_part_2** - 15 тысяч нулей из 40, может стоит сделать бинарным. Из выведенных признаков сильнее всего коррелирует с целевой переменной, скорее всего из-за числителя. (стало хуже на тысячные) (done: не добавлен)

6. **comm_part_3** - 15 тысяч нулей из 40, может стоит сделать бинарным. (done: не добавлен)