Закроют ли тему на StackOverflow

Кибитова Валерия

МГУ, ВМК, ММП

2015 г.

Постановка задачи

Задача:

Предсказать по заголовку, тексту темы и некоторым данным об авторе темы будет ли она закрыта

Критерий качества:

AUC-ROC

Исходные данные

- PostCreationDate
- OwnerUserId
- OwnerCreationDate
- ReputationAtPostCreation
- OwnerUndeletedAnswerCountAtPostTime
- Title
- BodyMarkdown
- Tag1, Tag2, Tag3, Tag4, Tag5

Извлечение дополнительных признаков

PostCreationDate, OwnerCreationDate

year, month, day, day_of_week, week, hour, minute, second

Title

length, words_count, parenthesis_count, ?_count, space_count

BodyMarkdown

length, words_count, parenthesis_count, ?_count, space_count, lines_count, tabs_count, brace_count, brackets_count, angle_brackets_count, =_count, @_count, dot_count, http_count

Tag1, Tag2, Tag3, Tag4, Tag5

tags_number, tagi_code, pair_features, counters

Метод локального контроля и Первые решения

Метод локального контроля

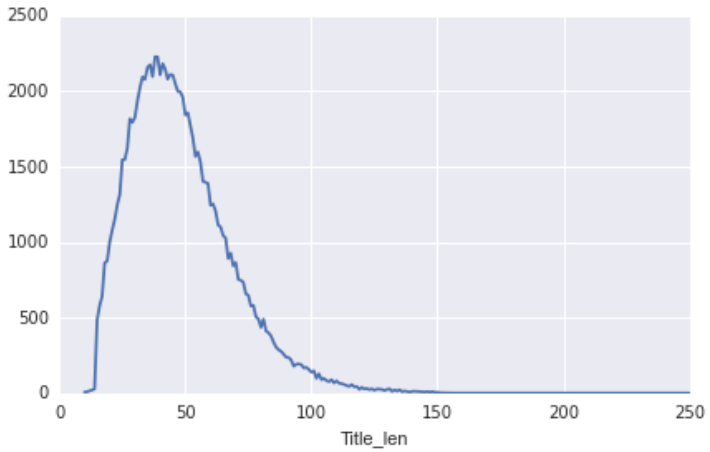
3 различных разделения на test и train, среднее значение auc по 3 тестам.

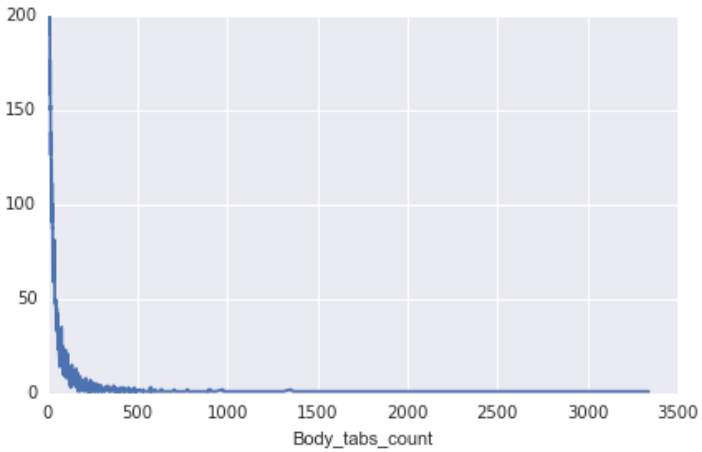
Исходное решение

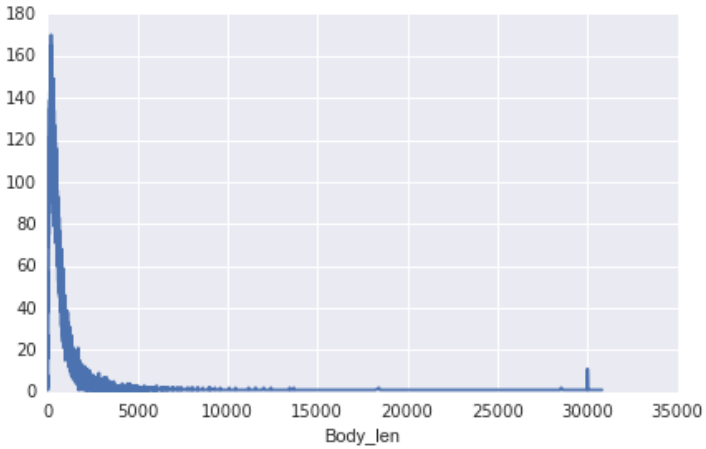
- Логистическая регрессия с подбором оптимального C на некоторых из начальных и дополнительных признаках(признаки выбирались путем изъятия признаков из полного набора) + TFidf отдельно над BodyMarkdown и отдельно над Title(качество на локальном контроле 0.8597)
- Случайный лес над другим подмножеством из тех же признаков(n_estimators=500, criterion ="entropy")(качество на локальном контроле 0.8076)
- Ансамбль из двух алгоритмов с подбором оптимальных коэффициентов(качество на локальном контроле: 0.87196)

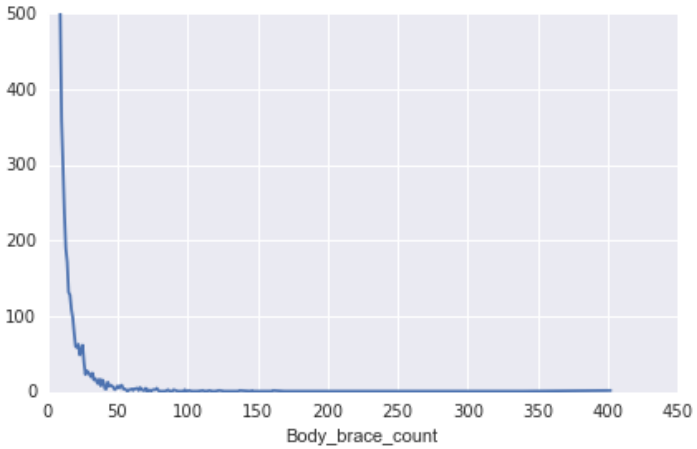
Логарифмирование признаков. Распределение признаков

Как можно заметить признаки распределены очень неравномерно.









Улучшения исходного алгоритма.

- Логарифмирование некоторых признаков LR(0.8679), ансамбль(0.8745712)
- Добавление строки тэгов к расчету tfidf для разметки и заголовка, добавление 2-граммов, LR(0.87292), ансамбль (0.87594)
- Добавление фичи OwnerUserId и CountsUserId, LR(0.87444), RF(0.80953), ансамбль(0.87708).

Итоговое качество в LB: 0.91428.

Спасибо за внимание!