Техническое задание.

Предскажите, как много звездочек наберет статья, зная только ее текст и время публикации.

Группа проекта: Осина Анна; Пляскин Павел

Для проекта мы выбрали данные с платформы kaggle.com, так как данный сайт представляет большой выбор соревнований и наборов данных для использования и применения алгоритмов машинного обучения. Мы будем предсказывать популярность статьи на Хабре. Ознакомиться с соревнованием можно перейдя по ссылке - https://inclass.kaggle.com/c/howpop-habrahabr-favs-lognorm

Задача заключается в предсказании количества звездочек, которые наберет статья, опубликованная на Хабре, по ее содержанию и времени публикации. Количество звездочек – это количество пользователей, которые добавили данную статью в раздел «избранное». Другими словами, количество звездочек определяет популярность статьи.

Целевой переменной является *favs_lognorm*.

В качестве метрики популярности статьи для данного соревнования используют «долю статей за последний месяц, у которых количество звездочек меньше чем у текущей статьи».

Число объектов в обучающей и тестовой выборках: ((134137, 17), (3990, 9))

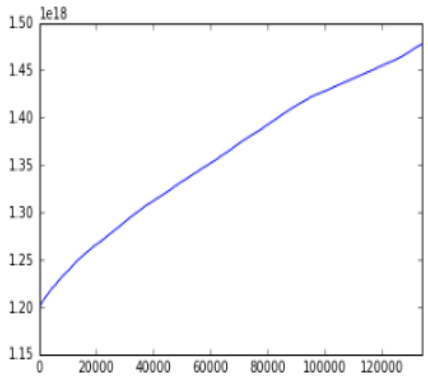
	0
url	https://habrahabr.ru/post/18284/
domain	habrahabr.ru
post_id	18284
published	2008-01-01 18:19:00
author	@Tapac
flow	develop
polling	False
content_len	4305
title	Новогодний подарок блоггерам — WordPress 2.3.2
comments	0
favs	0
views	236
votes_plus	0
votes_minus	0
views_lognorm	-0.792687
favs_lognorm	-1.34407
comments_lognorm	-2.43687

Признаки

url – единый указатель ресурса
domain – домен
post_id – номер поста
published – время публикации
author – автор
flow – тема статьи
polling – есть ли опрос в статье
content_len – длина контента
title – название
comments – комментарии
favs – избранное
views – просмотры
votes_plus – количество плюсов
votes_minus – количество минусов
views_lognorm – доля статей, у которых просмотров меньше, чем у текущей
favs_lognorm – доля статей, у которых звездочек меньше, чем у текущей
comments_lognorm – доля статей, у которых комментариев меньше, чем у текущей

Для объектов из тестовой выборки мы не будем знать значения никаких показателей популярности: *views, favs, comments, votes_plus, votes_minus, views_lognorm, comments_lognorm* и, соответственно, значение целевой переменной *favs_lognorm*.

```
train_df['published'].apply(lambda ts: pd.to_datetime(ts).value).plot();
```



Данные отсортированы по признаку *published*, что соответствует графику:

Метрикой для данного соревнования является MSE.

Для решения поставленной задачи предполагается:

- Провести предобработку данных (проверить наличие пропусков, выбросов и коррелирующих признаков, сделать их обработку, проанализировать данные на время наибольшего пика популярности статей, применить TfidfVectorizer, DictVectorizer, CountVectorizer)
- Обучить данные с помощью регрессионных моделей машинного обучения, а именно:
  - Linear Regression
  - Logistic Regression
  - Decision Tree Regression
  - Gradient Boosting regression
- Выбор наилучшей модели или ансамбля моделей