**2. Предсказание популярности новостной статьи**  
Датасет: OnlineNewsPopularity  
  
Описание: Датасет содержит информацию о статьях, опубликованных на веб-портале Mashable, и их популярности. Возможно, насколько популярна новостная статья можно предсказать по различным факторам, таким как количество слов, упоминание определенных ключевых слов, количество ссылок, количество комментариев, социальная активность на Facebook, Twitter и т. д.  
  
Предсказать популярность новостной статьи на основе заданных характеристик.  
  
*Данные:*

*В датасете представлены следующие признаки:*  
  
1. url - URL-адрес новостной статьи (строковый тип данных)  
2. timedelta - Количество дней между датой публикации статьи и 31 декабря 2014 года (вещественный тип данных)  
3. n\_tokens\_title - Количество слов в заголовке статьи (целочисленный тип данных)  
4. n\_tokens\_content - Количество слов в содержании статьи (целочисленный тип данных)  
5. n\_unique\_tokens - Отношение количества уникальных слов к общему числу слов в статье (вещественный тип данных)  
6. n\_non\_stop\_words - Отношение количества недоотстающих слов к общему числу слов в статье (вещественный тип данных)  
7. n\_non\_stop\_unique\_tokens - Отношение количества уникальных недоотдельных слов к общему количеству слов в статье (вещественный тип данных)  
8. num\_hrefs - Количество ссылок в статье (целое число)  
9. num\_self\_hrefs - Количество ссылок на статьи, опубликованные на том же веб-портале (целое число)  
10. num\_imgs - Количество изображений в статье (целочисленный тип данных)  
11. num\_videos - Количество встроенных видео в статье (целое число)  
12. average\_token\_length - Средняя длина слова в статье (вещественный тип данных)  
13. num\_keywords - Количество ключевых слов в статье (целое число)  
14. data\_channel\_is\_ ... - Бинарные признаки, показывающие категорию новостного канала (0 или 1)  
15. weekday\_is\_ ... - Бинарные признаки, отображающие день недели, в который была опубликована статья (0 или 1)  
16. global\_subjectivity - Cубъективность статьи, выраженная в количестве субъективных слов, относящихся к общему количеству слов статьи (вещественный тип данных)  
17. global\_sentiment\_polarity - Полярность статьи, выраженная через среднюю полярность всех фраз в статье (вещественный тип данных)  
18. global\_rate\_positive\_words - Отношение количества позитивных слов к общему числу слов в статье (вещественный тип данных)  
19. global\_rate\_negative\_words - Отношение количества негативных слов к общему числу слов в статье (вещественный тип данных)  
20. rate\_positive\_words - Отношение количества позитивных слов к общему числу слов в содержании статьи (вещественный тип данных)  
21. rate\_negative\_words - Отношение количества негативных слов к общему числу слов в содержании статьи (вещественный тип данных)  
22. avg\_positive\_polarity - Средняя положительность всех предложений в статье (вещественный тип данных)  
23. min\_positive\_polarity - Минимальная положительность предложений в статье (вещественный тип данных)  
24. max\_positive\_polarity - Максимальная положительность предложений в статье (вещественный тип данных)  
25. avg\_negative\_polarity - Средняя негативность всех предложений в статье (вещественный тип данных)  
26. min\_negative\_polarity - Минимальная отрицательность предложений в статье (вещественный тип данных)  
27. max\_negative\_polarity - Максимальная отрицательность предложений в статье (вещественный тип данных)  
28. title\_subjectivity - Субъективность заголовка статьи, выраженная в количестве субъективных слов, относящихся к общему количеству слов в заголовке статьи (вещественный тип данных)  
29. title\_sentiment\_polarity - Полярность заголовка статьи, выраженная через среднюю полярность всех фраз в заголовке статьи (вещественный тип данных)  
30. abs\_title\_subjectivity - Абсолютное значение относительной субъективности заголовка статьи (вещественный тип данных)  
31. abs\_title\_sentiment\_polarity - Абсолютное значение относительной полярности заголовка статьи (вещественный тип данных)  
32. shares - Число раз передачи статьи (популярность статьи)

Предварительная обработка данных:

1. Загрузка датасета и ознакомление с его структурой.

2. Удаление ненужных столбцов, таких как URL статьи или дата публикации, которые не влияют на популярность.

3. Обработка пропущенных значений, заполнение их средними или модами.

4. Преобразование категориальных переменных в числовые с помощью метода One-Hot Encoding или Label Encoding.

5. Масштабирование числовых признаков для более точной работы модели.

Визуализация данных:

6. Построение гистограммы распределения популярности статей для понимания их разброса.

7. Визуализация корреляции между признаками с помощью матрицы корреляции или тепловой карты.

8. Построение scatter plot для отображения зависимости популярности статей от выбранных признаков.

Построение регрессии:

9. Разделение данных на тренировочный и тестовый наборы.

10. Выбор и обучение модели регрессии, такой как линейная регрессия, случайный лес или градиентный бустинг.

11. Оценка точности модели с помощью метрик, таких как R2-коэффициент или среднеквадратическая ошибка.

Повторение шагов предварительной обработки и визуализации:

12. Оценка важности признаков с помощью feature importance для отбора наиболее значимых переменных.

13. Переобучение модели с использованием отобранных признаков.

14. Повторная оценка точности модели и сравнение с предыдущей моделью.

Тестирование модели:

15. Применение обученной модели к новым данным для предсказания популярности статей и оценка ее точности.