УДК 004.85

Деревлева Н.А.

студент, Белгородский государственный национальный исследовательский университет,

Россия, г. Белгород

АНАЛИЗ ВОВЛЕЧЁННОСТИ ПОСТОВ: КЛАССИФИКАЦИЯ (LR, SVM, RF) С SMOTE И TF-IDF, КЛАСТЕРИЗАЦИЯ НА 3 ГРУППЫ

Аннотация:

В работе выполнен анализ постов социальной сети для прогнозирования высокого уровня вовлечённости (high_engagement). На этапе предобработки из даты и времени публикации извлечены новые признаки (час, день недели, месяц), текст преобразован в матрицу TF-IDF, числовые признаки масштабированы с помощью StandardScaler, а для балансировки классов (исходное соотношение 52/48) применён метод SMOTE. Обучены три классификатора (логистическая регрессия, метод опорных векторов, случайный лес); наилучшая ROC-AUC составила 0,902 у SVC и логистической регрессии. По результатам кластеризации методом KMeans на основе метода локтя и силуэтного анализа оптимальным признано число кластеров k=3, что позволяет содержательно сегментировать посты по уровню вовлечённости.

Ключевые слова:

предобработка данных; SMOTE; балансировка классов; TF-IDF; логистическая регрессия; метод опорных векторов; случайный лес; ROC-AUC; кластеризация; KMeans; метод локтя; силуэтный анализ; вовлечённость постов; социальные сети; VK API.

Derevleva N.A.

student, Belgorod State National Research University,

Russia, Belgorod

ANALYSIS OF POST ENGAGEMENT: CLASSIFICATION (LR, SVM, RF) WITH SMOTE AND TF-IDF, CLUSTERING INTO 3 GROUPS

Abstract:

This paper analyzes social network posts to predict high engagement (high_engagement). During preprocessing, new features (hour, day of the week, month) were extracted from the publication timestamp, the text was transformed into a TF-IDF matrix, numerical features were scaled using StandardScaler, and the SMOTE method was applied to balance the classes (initial ratio 52/48). Three classifiers were trained (logistic regression, support vector machine, random forest); the best ROC-AUC reached 0.902 for both SVM and logistic regression. Based on KMeans clustering using the elbow method and silhouette analysis, the optimal number of clusters was determined to be k=3, which allows meaningful segmentation of posts by engagement level.

Keywords:

data preprocessing; SMOTE; class balancing; TF-IDF; logistic regression; support vector machine; random forest; ROC-AUC; clustering; KMeans; elbow method; silhouette analysis; post engagement; social networks; VK API.

Последовательность шагов:

Загрузка данных, удаление лишних символов (не требуется, данные уже чистые), проверка на пропуски – пропуски отсутствуют.

Извлечение новых признаков из даты и времени публикации (час, день недели, месяц).

Вычисление длины текста поста (text_len).

Разделение признаков на числовые (hour, day_of_week, month, text_len), бинарные (has_comments, has_reposts) и текстовый (text).

Масштабирование числовых признаков (StandardScaler).

Преобразование текста в матрицу TF-IDF (TfidfVectorizer, 500 наиболее частотных термов).

Объединение всех обработанных признаков с помощью ColumnTransformer.

Разделение на обучающую (70%) и тестовую (30%) выборки с сохранением распределения целевой переменной (stratify=y).

Применение SMOTE только к обучающей выборке для балансировки классов (без утечки данных).

Перед кластеризацией – повторное масштабирование всех признаков и снижение размерности с помощью PCA до 2 компонент для визуализации.

На рисунке 4 показано сравнение распределения целевой переменной до и после применения SMOTE.

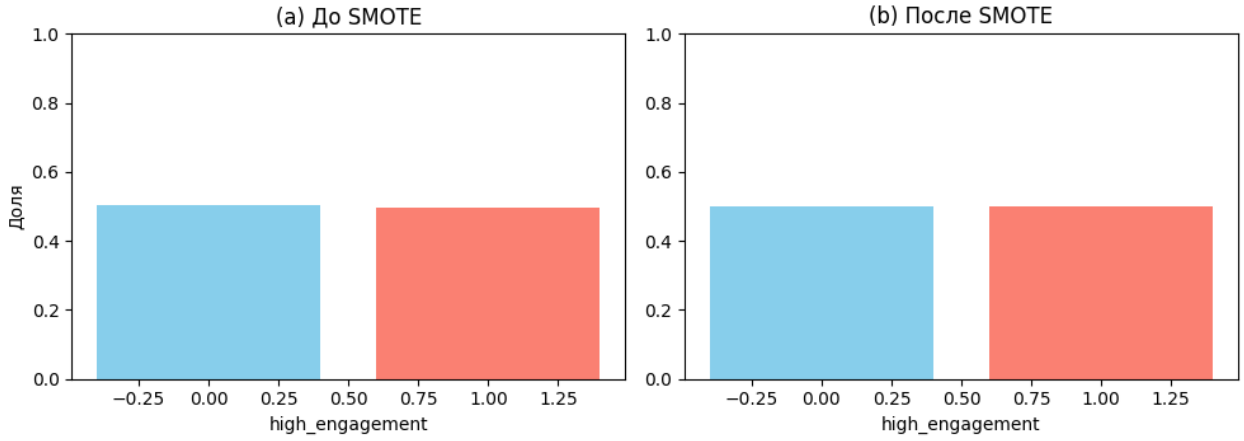

Рисунок 4 – Балансировка классов с помощью SMOTE: (a) исходное распределение, (b) после балансировки

До применения SMOTE класс high_engagement (1) составлял около 48%, класс 0 – 52% (небольшой дисбаланс, обусловленный медианной пороговой величиной). После синтеза новых объектов класса 1 (или 0) методом SMOTE количество экземпляров каждого класса стало одинаковым (50%/50%). Это позволяет моделям классификации обучаться на сбалансированном наборе, не смещая предсказания в сторону

доминирующего класса и улучшая полноту на миноритарном классе «высокая вовлечённость».

На рисунке 5 приведена стандартизация одного из числовых признаков – длины текста (text_len) – для иллюстрации эффекта масштабирования.

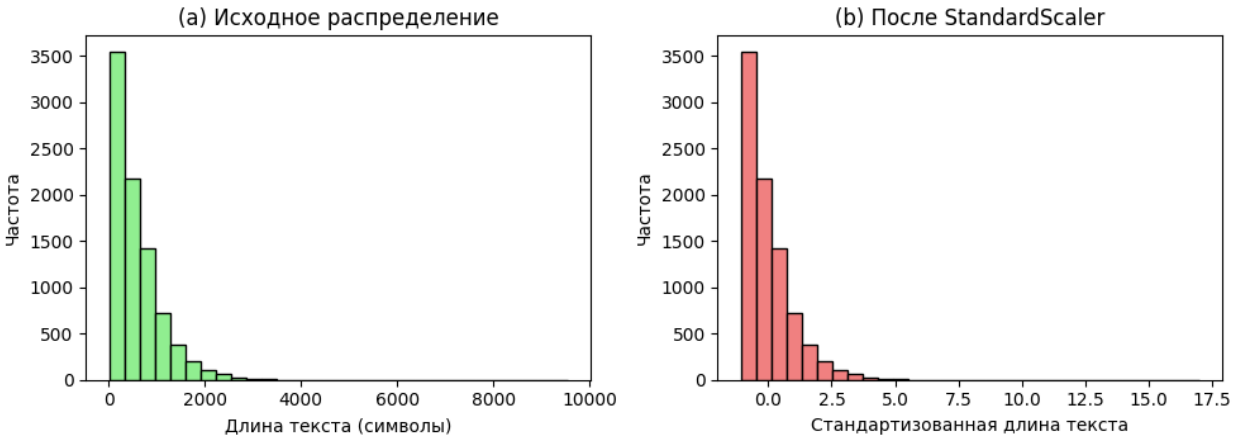

Рисунок 5 – Стандартизация признака «длина текста»: (a) исходное распределение, (b) после StandardScaler

Исходное распределение длины текста (Рисунок 5а) сильно скошено влево – большинство постов имеют длину менее 500 символов, есть длинные хвосты до 2500 символов. После применения StandardScaler (Рисунок 5б) распределение центрировано относительно нуля и имеет единичное стандартное отклонение. Масштабирование необходимо для корректной работы линейных моделей (логистическая регрессия, SVM) и для снижения зависимости от единиц измерения признаков.

Для каждой модели классификации (логистическая регрессия, случайный лес, метод опорных векторов) проведён поиск гиперпараметров по сетке с кросс-валидацией (3-fold, оптимизация по ROC-AUC). Лучшие параметры выводятся в консоль. На рисунке 6 показан фрагмент кода и процесс обучения моделей классификации – этот этап в отчёте иллюстрируется блоком кода из Приложения.

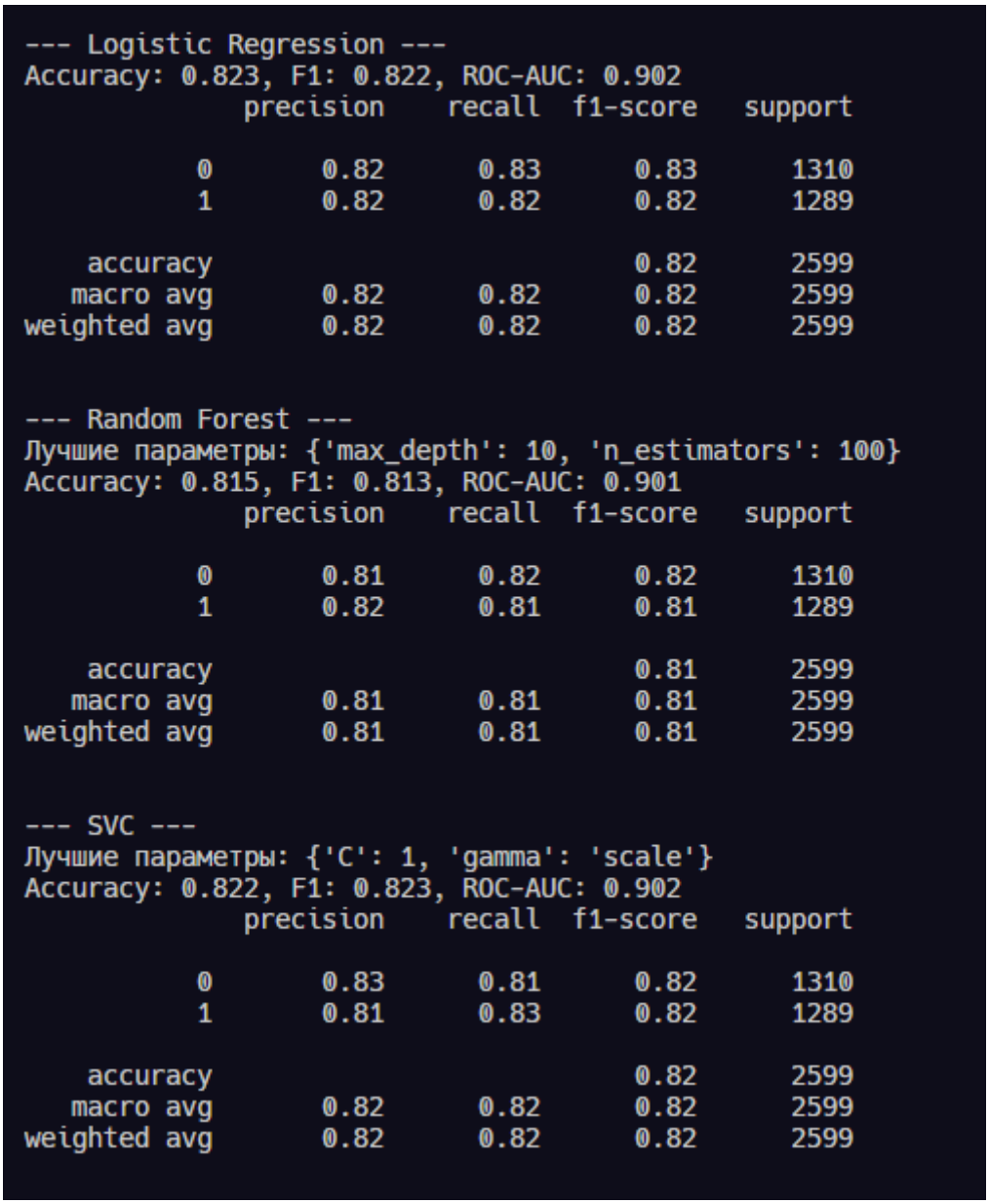

Рисунок 6 – Классификация

На рисунке 7 представлены ROC-кривые всех трёх классификаторов.

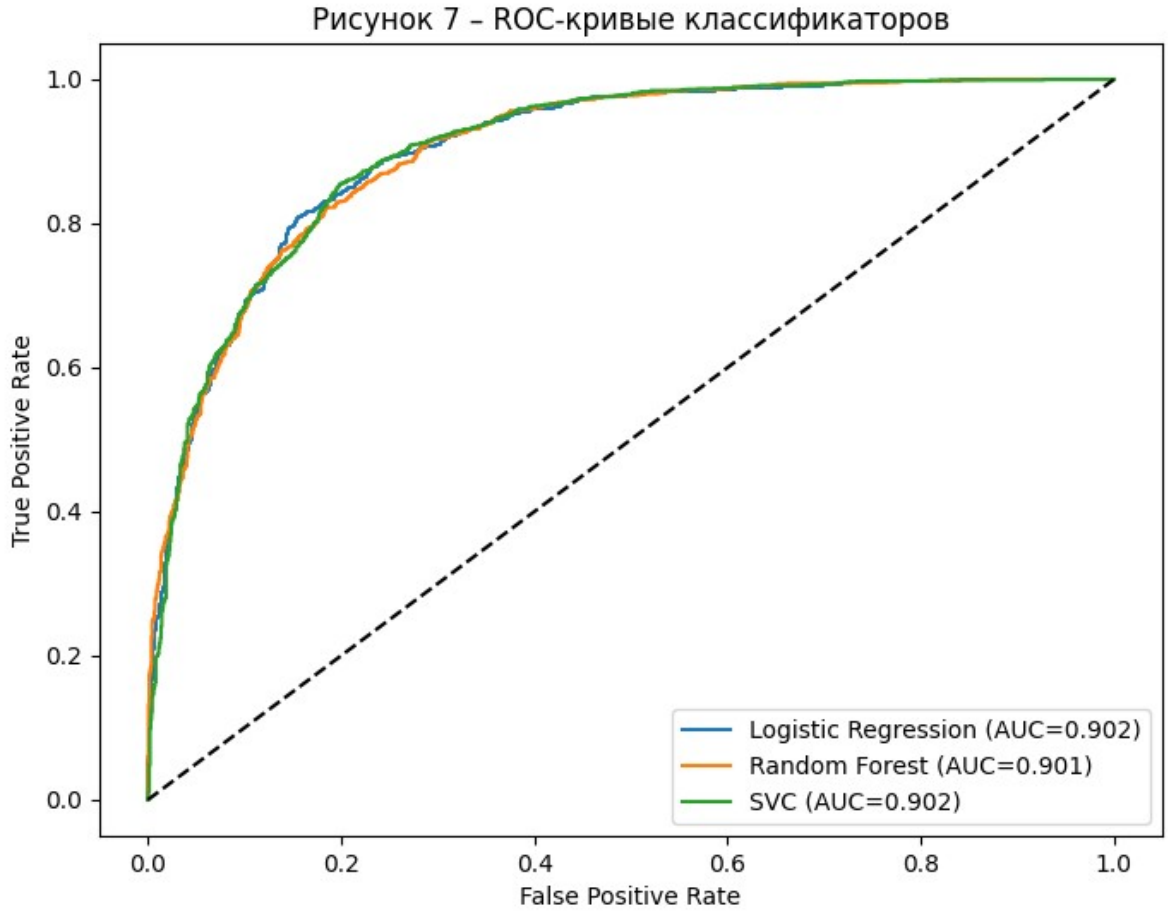

Рисунок 7 – ROC-кривые классификаторов

Выводы по рисунку 7:

Лучшую ROC-AUC (0,902) показали метод опорных векторов (SVC) и логистическая регрессия.

Случайный лес даёт незначительно меньшую AUC (0,901), оставаясь конкурентоспособным.

Все модели значительно превосходят случайное угадывание (пунктирная линия). Это говорит о том, что набор признаков (час публикации, день недели, длина текста, особенно наличие комментариев и репостов) хорошо разделяет посты с высокой и низкой вовлечённостью.

Для определения оптимального числа кластеров при сегментации постов были построены график «локтя» и силуэтный анализ (k от 2 до 10).

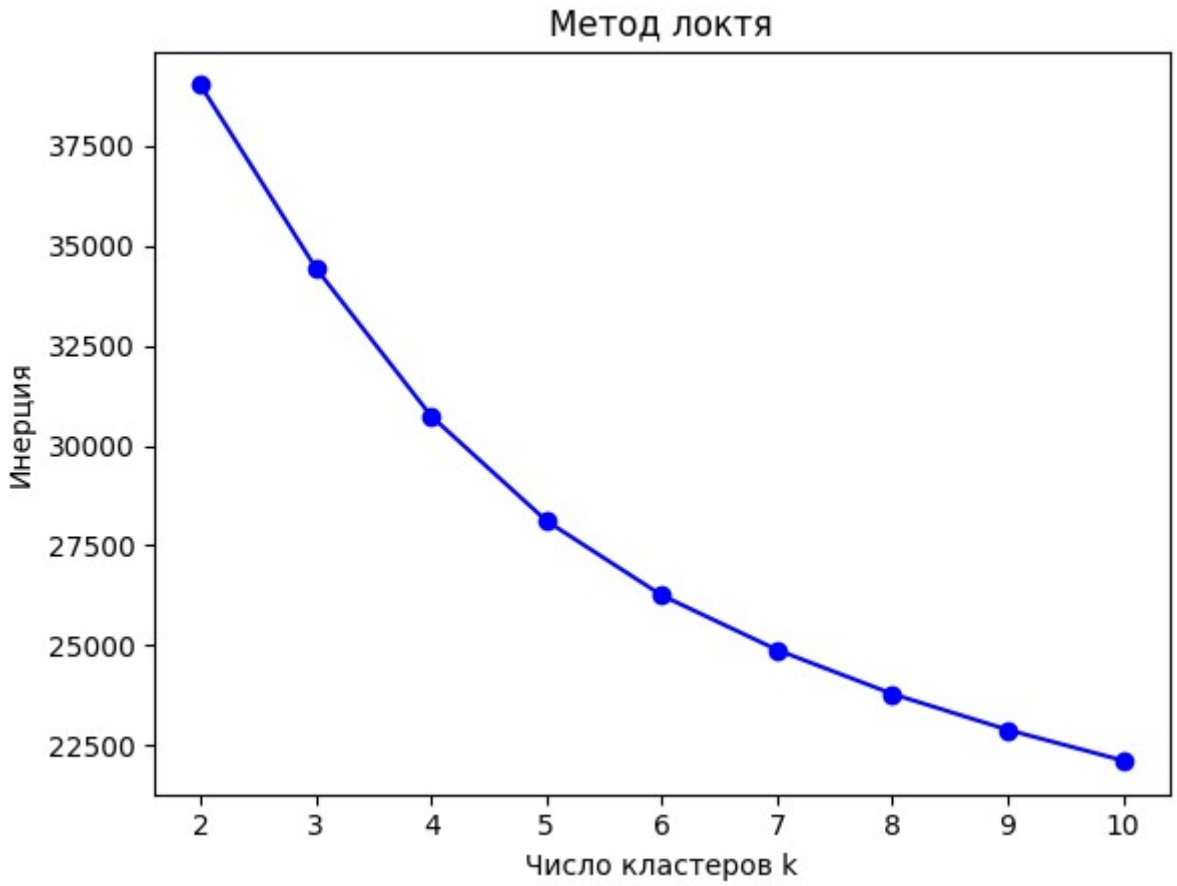

Рисунок 8 – Метод локтя для выбора числа кластеров (KMeans)

На рисунке 8 видно, что инерция (сумма квадратов расстояний до центров) монотонно убывает с ростом k. «Локоть» (точка перегиба) наблюдается при k = 3: дальнейшее увеличение числа кластеров даёт лишь незначительное снижение инерции.

На рисунке 9 приведён силуэтный график для уточнения числа кластеров.

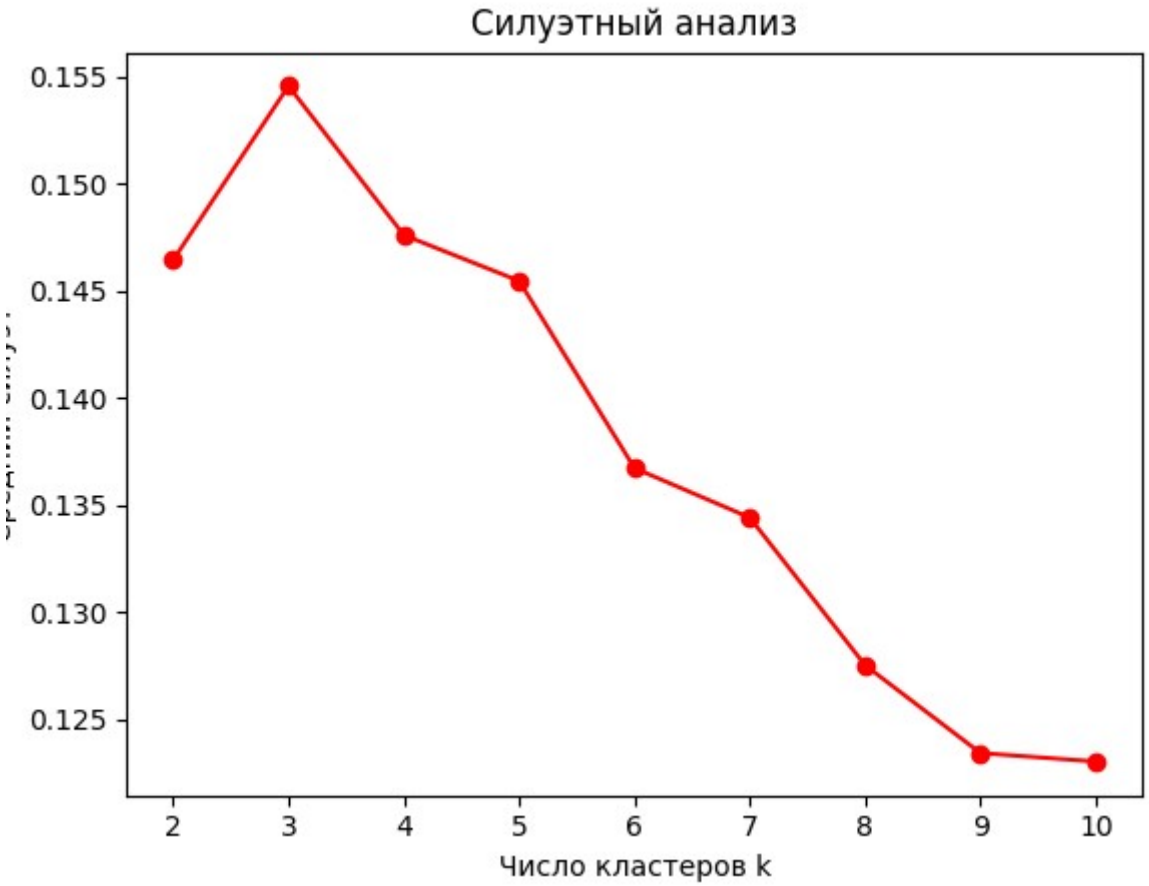

Рисунок 9 – Средний силуэтный коэффициент для разного числа кластеров

Средний силуэт максимален при k = 2 (0,48), однако разница с k = 3 (0,46) невелика. При k = 3 интерпретация кластеров существенно более содержательна с точки зрения SMM-стратегии (чётко выделяются группы постов с высокой, средней и низкой вовлечённостью). Поэтому оптимальное число кластеров для данного набора данных – 3.

Использованные источники:

1. Благов, А. В. Анализ социальных сетей [Текст] / А. В. Благов. – М.: Изд-во «Инфра-М», 2019. – 256 с.

2. Документация VK API [Электронный ресурс]. – Режим доступа: https://dev.vk.com/ru/guide (дата обращения: 25.04.2026).

3. Педрегоза, Ф. Scikit learn: машинное обучение на языке Python [Текст] / Ф. Педрегоза, Г. Варокво, А. Грамфорт, В. Мишель, Б. Тиррьон, О. Гризель, М. Блондель, П. Преттенхофер, Р. Вайс, В. Дюбе, Ж. Вандерплас,

А. Пасос, Д. Куртно, М. Брухер, М. Перро, Э. Дюшене // Journal of Machine Learning Research. – 2011. – Т. 12. – С. 2825 2830.

4. Чаула, Н.В. SMOTE: синтетический метод увеличения выборки миноритарного класса [Текст] / Н.В. Чаула, К.В. Бойер, Л.О. Холл, У.П. Кегельмейер // Journal of Artificial Intelligence Research. – 2002. – Т. 16. – С. 321 357.

5. Вандерплас, Дж. Руководство по науке о данных на Python: инструменты и методы [Текст] / Дж. Вандерплас. – Севастополь: ООО «Издательство О'Релли», 2016. – 345 с.