МЕТОДЫ РЕШЕНИЯ ПРОБЛЕМЫ НЕСБАЛАНСИРОВАННОСТИ НАБОРА ДАННЫХ ПРИ ВЫЯВЛЕНИИ ВРЕДОНОСНЫХ ВЕБ-АДРЕСОВ

Кабулов Анвар Василович

Профессор Национального университета Узбекистана имени Мирзо Улугбека

E-mail: anvarkabulov@mail.ru

Отахонов Алишербек Абдумутал угли

Базовый докторант Ферганского государственного университета

E-mail: alisherotaxonov91@gmail.com

Аннотация: несбалансированные наборы данных представляют собой серьезную проблему в области обнаружения вредоносных веб-адресов. Неравномерное распределение образцов вредоносных и безвредных веб-адресов представляет собой проблему для современных систем обнаружения на основе машинного обучения, поскольку создает неточные модели и приводит к низкому уровню обнаружения вредоносных веб-адресов. В этом исследовании изучается эффективность классификаторов машинного обучения (ML) на наборах данных с высоким уровнем неравенства с использованием двух методов повышения частоты дискретизации — SMOTE и ADASYN — и проводится комплексная оценка эффективности классификатора с использованием оценочных показателей. Исследование показало, что метод выборки SMOTE дал наиболее эффективные результаты оценки при использовании со случайным лесом (RF). На основе выбранной наилучшей комбинации прогнозируется новая база данных для фишингового адреса.

Ключевые слова: машинное обучение, несбалансированный набор данных, избыточная выборка, SMOTE, ADASYN, Случайный лес.

Annotatsiya: balanssiz ma'lumotlar to'plamlari zararli veb manzillarni aniqlash sohasida jiddiy muammo hisoblanadi. Zararli va zararsiz veb manzillar namunalarining notekis taqsimlanishi zamonaviy mashina o'rganishga asoslangan aniqlash tizimlari uchun qiyinchilik tug'diradi, chunki u noto'g'ri modellarni yaratadi va zararli veb manzillarni aniqlashning past tezligini keltirib chiqaradi. Ushbu tadqiqot mashinali o‘qitish (Machine Learning – ML) klassifikatorlarining samaradorligini ikkita yuqori namuna olish texnikasi - SMOTE, ADASYN usullarini qo’llab nomutanosiblik darajalariga ega bo'lgan ma'lumotlar to'plamlari bo'yicha o'rganiladi va baholash ko'rsatkichlaridan foydalangan holda tasniflagichning ishlashini har tomonlama baholanadi. Tadqiqot natijasida SMOTE namuna olish usuli Random Forest (RF) bilan foydalanilganda eng samarali

ISSN 3030-3907

baholash natijasini berdi. Tanlab olingan eng yaxshi kombinatsiya asosida yangi ma’lumotlar bazasini fishing manzilga bashorat qiladi.

Kalit so'zlar - mashinani o'rganish; nomutanosiblik ma'lumotlar to'plami; ortiqcha namuna olish; SMOTE; ADASYN, Random Forest.

Abstract: imbalanced datasets are a serious problem in the field of malicious web address detection. The uneven distribution of malicious and non-malicious web address samples poses a challenge for modern machine learning-based detection systems, as it creates incorrect models and leads to a low rate of malicious web address detection. This study investigates the performance of machine learning (ML) classifiers on datasets with high levels of imbalance using two sampling techniques - SMOTE, ADASYN methods, and comprehensively evaluates the performance of the classifier using evaluation metrics. As a result of the study, the SMOTE sampling method gave the most effective evaluation result when used with Random Forest (RF). Based on the best combination selected, it predicts a new database for a phishing address.

Keywords: machine learning; imbalance datasets; oversampling; SMOTE; ADASYN, Random Forest.

Введение

Кибербезопасность является важной областью в современном мире, и обнаружение вредоносных веб-адресов является важной областью кибербезопасности, поскольку количество вредоносных веб-адресов быстро увеличивается и становится все более сложным [1]. Более того, решения для обнаружения вредоносных веб-адресов очень важны, учитывая острую необходимость решения проблемы. Однако из-за сложности современных вредоносных веб-адресов и объема генерируемых данных стало сложно более эффективно выявлять вредоносные веб-адреса. В ответ на эту проблему методы машинного обучения (Machine Learning – ML) приобрели популярность, изучая вредоносные веб-адреса и отличая их от безвредных веб-адресов [2]. Однако проблема несбалансированности наборов данных является серьезным препятствием для разработки эффективных систем обнаружения вредоносных веб-адресов. Это происходит из-за того, что набор данных, используемый для обучения моделей ML, содержит гораздо меньше вредоносных образцов, чем данные, связанные с безвредными образцами, что приводит к неточным моделям, которые не могут адекватно обнаружить вредоносную активность.

Постановка проблемы

В данной статье мы представляем методы решения проблемы несбалансированности набора данных при выявлении вредоносных веб-

ISSN 3030-3907

адресов и оцениваем их эффективность посредством систематического обзора. На основании результатов оценки мы проверяем и прогнозируем, являются ли новые веб-адреса вредоносными или легитимными.

ЛИТЕРАТУРНЫЙ ОБЗОР

Несбалансированный набор данных для выявления вредоносных веб-адресов

Наборы данных для обнаружения вредоносных веб-адресов несбалансированы, когда образцы вредоносных веб-адресов (класс меньшинства) имеют очень небольшое распределение по сравнению с невредоносными образцами (класс большинства) [2].

В результате прогнозы моделей, как правило, искажают класс большинства и часто игнорируют важные модели меньшинства, которые являются основным направлением кибербезопасности. На рисунке 1 показано непропорциональное распределение данных.

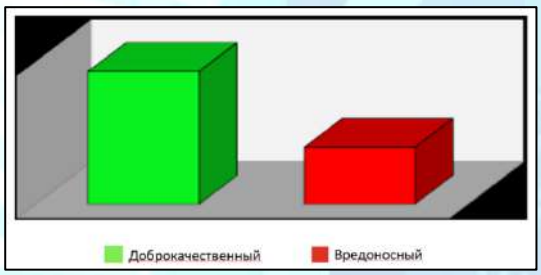


Рисунок 1. Визуальное представление несбалансированного набора данных.

Сбалансированный набор данных для выявления вредоносных веб-адресов

Сбалансированные наборы данных для обнаружения вредоносных веб-адресов — это наборы данных, которые имеют примерно одинаковое количество образцов в классе большинства (невредоносная информация) и классе меньшинства (вредоносная информация) [2]. Этот баланс предотвращает смещение классификаторов в сторону какого-либо одного класса и приводит к более точной идентификации незараженных и вредоносных данных. На рисунке 2 показано сбалансированное распределение данных, при котором оба данных вычислительно равны.

ISSN 3030-3907

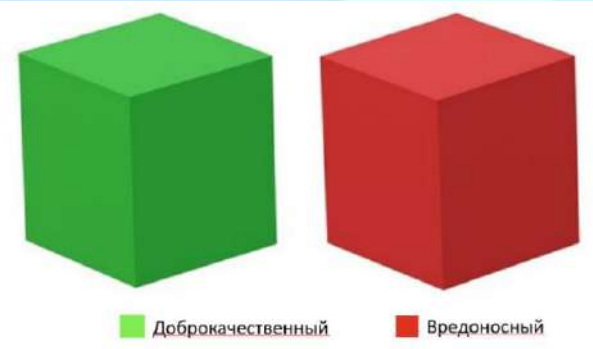


Рисунок 2. Визуальное представление сбалансированного набора данных.

Подход на уровне данных для решения несбалансированных наборов данных

Существуют различные методы решения проблемы несбалансированности наборов данных при выявлении вредоносных веб-адресов. Мы в целом классифицируем один из этих подходов как методы на уровне данных, решающие проблему дисбаланса.

Подходы на уровне данных пытаются сбалансировать распределение классов путем преобразования данных перед применением любого алгоритма ML [2]. В таблице 1 показан метод на уровне данных.

Таблица 1.

Обзор методов на уровне данных

Метод	Определение	Преимущества	Ограничения
Передискретизация [4]	Добавляет синтетические примеры в класс меньшинства, чтобы сбалансировать набор данных.	Балансирует распределение классов без потери существующих данных.	Риск переобучения и вычислительных затрат при обработке больших синтетических наборов данных [6].
Недостаточная дискретизация [6]	Большинство сокращают выборки классов, чтобы сбалансировать набор данных, путем случайного удаления примеров или использования	Упрощает набор данных и заставляет модель одинаково фокусироваться на обоих классах.	Потеря потенциально ценной информации от класса большинства.

ISSN 3030-3907

	эвристических методов.		

На рисунке 3 показано визуальное представление избыточной и недостаточной выборки. Это наглядно иллюстрирует разницу между методами OverSampling и UnderSampling.

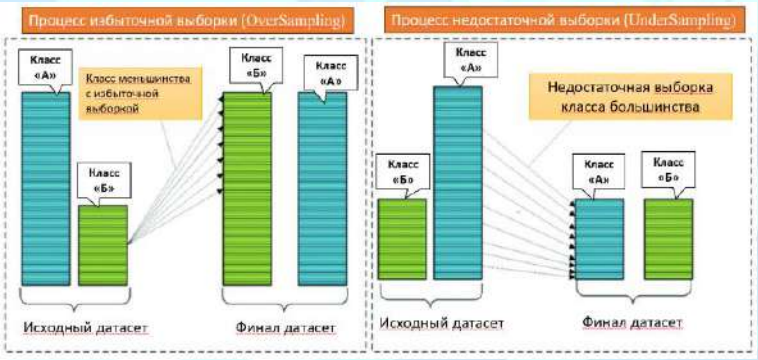


Рисунок 3. Иллюстрация методов OverSampling и UnderSampling для обработки несбалансированных наборов данных.

Несбалансированные методы балансировки данных

Несбалансированные наборы данных являются распространенной проблемой при интеллектуальном анализе данных. В нашем исследовании мы проанализировали наборы данных с различной степенью неравенства, которые (постоянно) показывают доминирующий класс большинства по отношению к классу меньшинства. В наборе данных (DS1) различия варьируются от 83,2% до 16,8%. База данных (DS2), используемая при прогнозировании модели, была выбрана в соотношении 83,1% к 16,9%.

Мы предоставляем следующую информацию о методах балансировки SMOTE и ADASYN, которые используются в этой статье.

Метод передискретизации SMOTE

SMOTE (Synthetic Minority Oversampling Technique – метод синтетической перевыборки меньшинств) — популярный метод, разработанный для решения проблемы несбалансированных наборов данных, которая является распространенной проблемой в области интеллектуального анализа данных и машинного обучения [3]. SMOTE улучшает баланс набора данных, генерируя синтетические примеры для класса меньшинства вместо дублирования существующих.

Последующие исследования, основанные на основополагающей работе SMOTE, ввели множество модификаций и адаптаций, таких как Borderline-SMOTE [4], Safe-Level-SMOTE и ADASYN [5], для решения проблем с данными и устранения особенностей.

Метод передискретизации ADASYN

ISSN 3030-3907

ADASYN (адаптивная синтетическая выборка) представляет собой значительный прогресс в решении проблемы дисбаланса классов в машинном обучении и анализе данных [5].

ADASYN работает на основе механизма взвешивания, который определяет количество синтетических образцов на основе сложности обучения для каждого меньшинства образцов. Этот механизм определяется плотностью большинства примеров в меньшинстве. Чем больше выборка меньшинства окружена выборками большинства, тем больше создается синтетических выборок. Порог принятия решения ADASYN фокусируется на более сложных областях, чтобы значительно улучшить соответствие меньшинства.

Показатели оценки

Для оценки моделей МО используется оценка на основе индикаторов. Целью данной оценки является минимизация вероятности ошибок путем количественной оценки различий между прогнозируемыми и фактическими баллами класса. Эти показатели имеют решающее значение для определения того, делает ли классификатор правильные прогнозы в пределах заданного порога принятия решения.

Для оценки моделей МО, используемых для обнаружения вредоносных веб-адресов, необходимы такие методы оценки, как полнота, точность и оценка F1 [6]. Матрица путаницы, в которой подробно описаны истинно положительные (TP), ложноположительные (FP), ложноотрицательные (FN) и истинно отрицательные (TN) результаты, предоставляет важную информацию о точности оценки между классами.

Используя значения из Таблицы 2, результаты точности (1), достоверности (2) и полноты (3) позволяют провести комплексную оценку эффективности модели. Формула 1 измеряет качество положительных предсказаний класса, Формула 2 оценивает и отражает общую точность предсказаний, а Формула 3 оценивает и отражает их полноту:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

ISSN 3030-3907

Таблица 2.

Матрица путаницы

		Прогнозирующий класс	
		Фишинг	Реальный
Реальный класс	Фишинг	TP	FN
	Реальный	FP	TN

Оценка F1 объединяет точность и полноту в одну метрику, предлагая сбалансированную метрику производительности, которая учитывает оба аспекта:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Чем ближе значение F1-оценки к 1, тем лучше модель достигает баланса между точностью и полнотой, что отражает более высокую общую производительность [7].

Площадь под кривой ROC (ROC AUC) является ключевым показателем при прогнозировании отключений электроэнергии. Кривая ROC показывает компромисс между истинно положительным показателем (TPR или полнота) и ложноположительным показателем (FPR) при различных пороговых значениях классификации при оценке способности классификатора различать классы:

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

Наше исследование соответствует этим передовым практикам, уделяя особое внимание эффективности модели в управлении дисбалансом классов, а не оптимизации большинства прогнозов классов при оценке вредоносных веб-адресов. Точность и полнота также являются исключительными независимыми показателями, поскольку оценка F1 адекватно отражает их совокупный эффект. Эта система оценки, включающая оценку F1, ROC AUC и точность, обеспечивает комплексную оценку общей точности модели и ее способности обнаруживать фишинговые веб-страницы в несбалансированных наборах данных.

Датасеты

Описание данных

В рамках данного исследования были использованы два набора данных из общедоступных источников, далее именуемые DS1 и DS2. Оба набора данных взяты с платформы Kaggle. Из полученного набора данных с

ISSN 3030-3907

помощью программного обеспечения было извлечено 29 признаков для создания нового набора данных. Оба набора данных содержат 6000 записей. Набор данных включает в себя такие характеристики, как длина URL, содержимое домена (например, количество точек, количество TLD, количество специальных символов), параметры URL и длина части файла, а также количество символов в них [7]. Распределение классов в этом наборе данных показывает соотношение 1:6, при этом строки, помеченные как фишинг, составляют примерно 16% данных. Первая из этих баз данных (DS1) используется в обучающей и оценочной части модели. Мы используем вторую базу данных для составления прогнозов с использованием наилучшей комбинации. Информация о строках в используемых нами базах данных показана на рисунке 4.

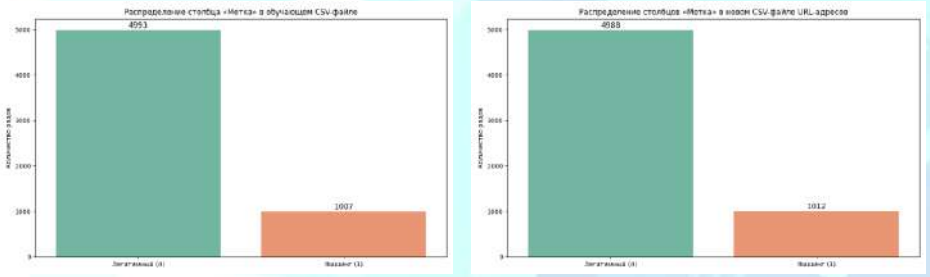


Рисунок 4. Содержимое баз данных, использованных в статье.

ПРЕДЛАГАЕМАЯ МЕТОДОЛОГИЯ

Методология исследования систематически разделена на семь отдельных этапов: загрузка и обработка набора данных, извлечение признаков, генерация несбалансированных наборов данных, применение методов выбора и выборки модели, обучение модели, оценка модели и прогнозирование.

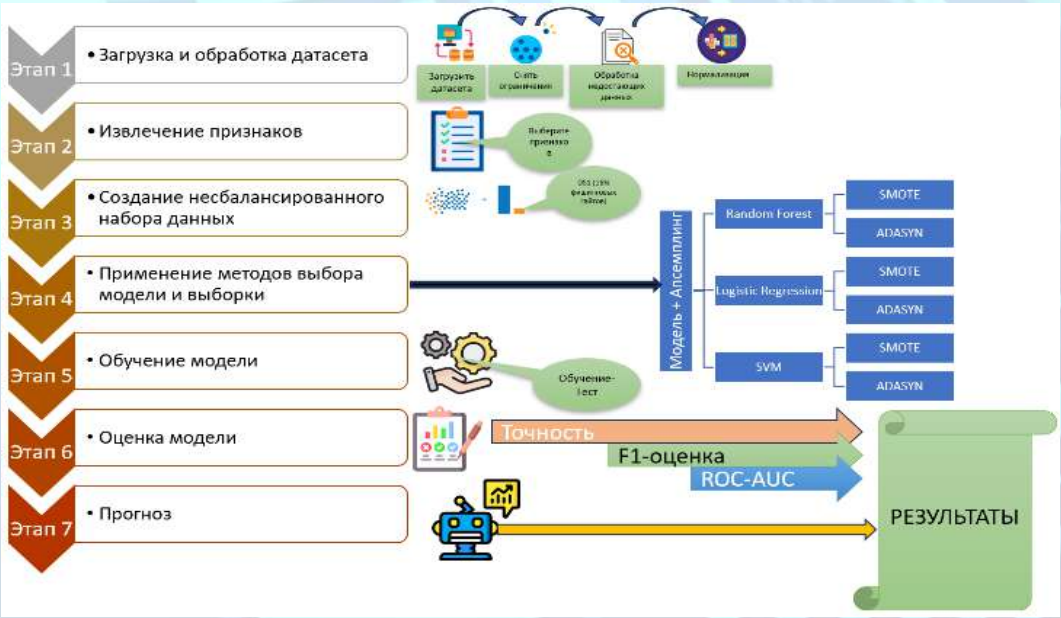


Рисунок 5. Этапы методологии исследования.

ISSN 3030-3907

В данном исследовании рассматривается вопрос прогнозирования веб-адресов для фишинговых угроз. В нем рассматривается, как методы избыточной выборки, такие как SMOTE и ADASYN, повышают точность оценки этих моделей. Для получения надежных и обобщаемых результатов был проанализирован большой объем данных.

Разделение данных

Наборы данных были разделены на обучающие и тестовые подмножества, при этом 70% было выделено на обучение и 30% на тестирование. Стратифицированное разбиение гарантирует, что пропорции распределения классов в исходном наборе данных сохраняются в обоих подмножествах. Такой подход имеет решающее значение для поддержания баланса классов, что позволяет обучать и оценивать модель на данных, которые точно отражают распределение набора данных.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результаты DS1 показывают 16% уровень фишинга

Оценка показывает, что модель RF последовательно превосходит модели LR и SVM при использовании с методом выборки SMOTE. В рамках SMOTE случайный лес достигает наивысших результатов по всем ключевым показателям, в том числе на 4% и 5% выше, чем модели LR и SVM по точности, на 11% и 19% выше по F1-оценке и на 7% и 5% выше по значению ROC-AUC соответственно. Это показывает тесную связь между расширенными возможностями обучения RF и генерацией синтетических образцов SMOTE.

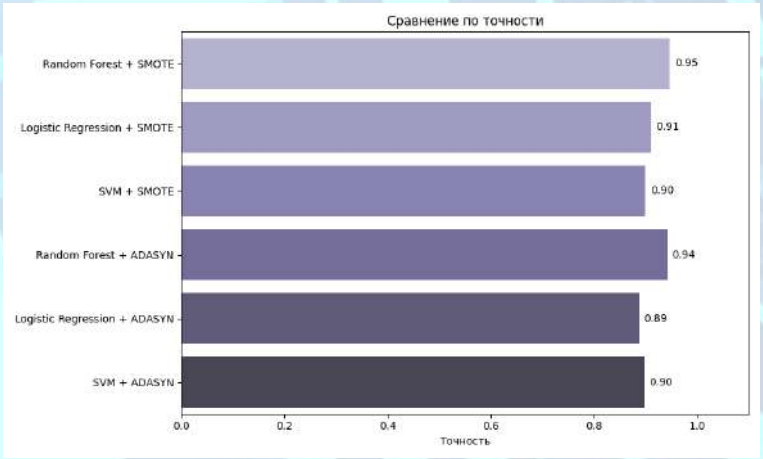


Рисунок 6. Значения точности для DS1 комбинации моделей и методов

Даже при использовании метода ADASYN модель RF показала лучшие и более точные результаты, чем другие модели. То есть, ее точность составила 94%, тогда как у остальных моделей она составила 89% и 90% соответственно. Модель RF также показала более высокий результат по F1-

ISSN 3030-3907

оценке. Модель RF (97%), LR (87%) и SVM (92%) также показали хорошие результаты с точки зрения оценки ROC-AUC.

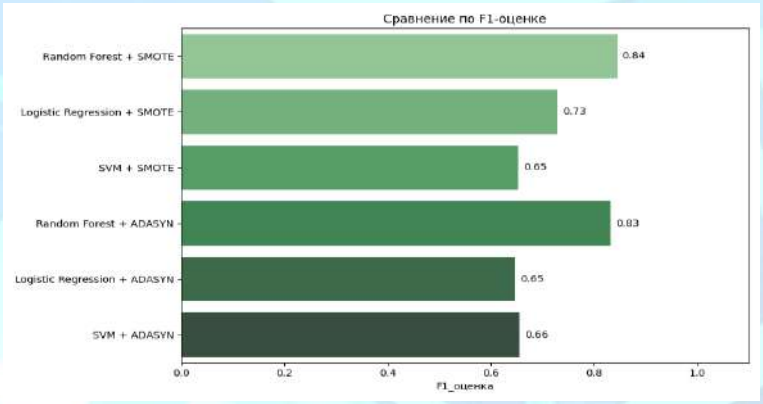


Рисунок 7. Значения F1-оценки для DS1 комбинации моделей и методов

Таблица 3.

Жирным шрифтом выделены оценки моделей для DS1 с коэффициентом фишинговой линии 16%, а звездочками отмечены наибольшие значения.

Модели /Оценки	Точность (%)	F1-оценка (%)	ROC-AUC (%)
RF+SMOTE	**95***	**84***	**97***
LR+SMOTE	91	73	90
SVM+SMOTE	90	65	92
RF+ADASYN	94	83	**97***
LR+ADASYN	89	65	87
SVM+ADASYN	90	66	92

Матрица путаницы комбинации моделей и методов, которые мы исследуем, показана на рисунке 8 ниже.

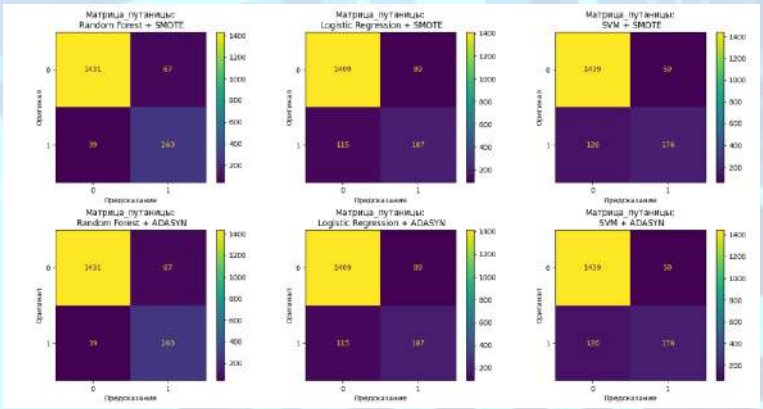


Рисунок 8. Матрица путаницы для комбинаций моделей и методов.

ISSN 3030-3907

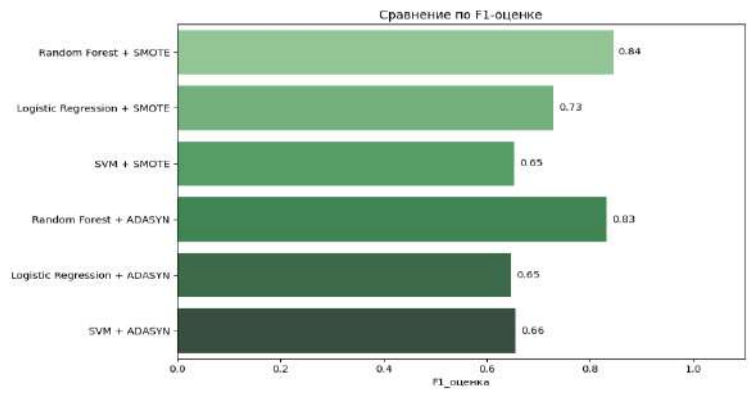


Рисунок 9. Различные комбинации моделей для оценки ROC-AUC и методов выборки с 16%-ным соотношением строк фишинговых веб-адресов.

На основании вышеприведенного анализа можно сделать вывод, что модель Random Forest+SMOTE можно считать наиболее эффективной улучшенной моделью для проверки веб-адресов на предмет фишинга.

В целом, алгоритм RF, улучшенный с помощью SMOTE, значительно превосходит ADASYN в устранении дисбаланса классов, обеспечивая лучшую производительность. Результаты, представленные в таблице 3 и на рисунках 6–9, подтверждают, что SMOTE является наиболее эффективным методом, а Random Forest — наиболее надежной моделью для этого набора данных.

На основании приведенных выше результатов мы выбрали комбинацию Random Forest + SMOTE как наиболее эффективную модель. Используя эту комбинацию, мы прогнозируем, что база данных DS2 является фишинговым адресом. Результат прогноза показан на рисунке 10 ниже. По его словам, выбранная модель дала прогноз с точностью 88,6%. Видно, что выбранная модель более полезна в случае непропорционального фишингового веб-адреса.

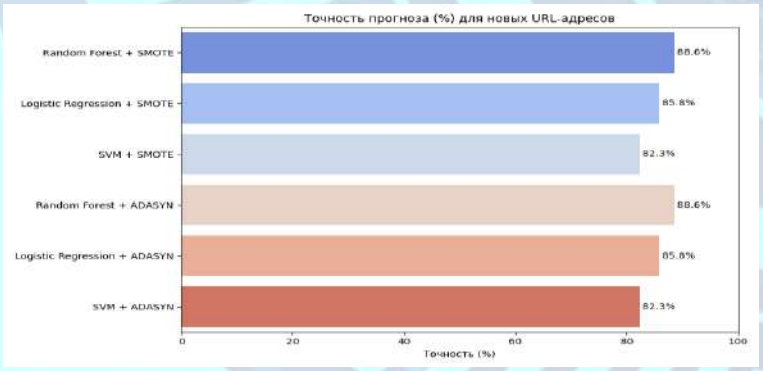


Рисунок 10. Результат прогнозирования DS2 на фишинговый адрес на основе комбинации Random Forest+SMOTE

ЗАКЛЮЧЕНИЕ

ISSN 3030-3907

В данном исследовании изучалось взаимодействие между контролируемыми моделями машинного обучения — случайным лесом, логистической регрессией и SVM — и методами выборки SMOTE, ADASYN. Было оценено их совокупное влияние на эффективность прогнозирования на наборе данных с дисбалансом, составляющим 16% фишинговой строки. Основная цель состояла в том, чтобы определить, какая комбинация модели и метода выборки обеспечит наиболее точные прогнозы в умеренно искаженном наборе данных.

В этом исследовании подчеркивается важность выбора подходящей комбинации моделей машинного обучения и методов выборки для решения проблем несбалансированных наборов данных при прогнозировании фишинговых веб-адресов. Результаты показывают, что случайный лес в сочетании с SMOTE оказывается наиболее эффективным решением, когда уровень фишинговых строк составляет 16% для умеренно несбалансированных наборов данных. Такое сочетание обеспечивает превосходную точность оценки. Точность составила 94%, оценка F1 — 84%, а показатель ROC-AUC — 97%. Эта комбинация спрогнозировала фишинговый адрес с точностью 88,6% при применении к новой базе данных URL-адресов. Эти результаты подчеркивают надежность и устойчивость комбинации Random Forest + SMOTE при обработке несбалансированных данных.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

1. A.Alharbi, A. H. Seh, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, “Analyzing the impact of cyber security related attributes for intrusion detection systems,” *Sustainability*, vol. 13, no. 22, p. 12337, 2021.

2. H. Ali, M. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, “Imbalance class problems in data mining: A review,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1560–1571, 2019.

3. Hemmatian, J., Hajizadeh, R., & Nazari, F. (2025). Addressing imbalanced data classification with cluster-based reduced noise SMOTE. *PLoS One*, *20*(2) doi:https://doi.org/10.1371/journal.pone.0317396

4. Ahsan, M. M., Raman, S., & Siddique, Z. (2023). *BSGAN: A novel oversampling technique for imbalanced pattern recognitions*. Ithaca: Retrieved from https://www.proquest.com/working-papers/bsgan-novel-oversampling-technique-imbalanced/docview/2815043632/se-2

5. Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Alvarado, E. S., Isabel, D. L. T. D., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: An

ISSN 3030-3907

investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(1), 87. doi:https://doi.org/10.1186/s40537-024-00943-4

6. Owusu-Adjei, M., James Ben Hayfron-Acquah, Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, *2*(11) doi:https://doi.org/10.1371/journal.pdig.0000290

7. А А. Отахонов ОБНАРУЖЕНИЕ И ОЦЕНКА ФИШИНГОВЫХ URL-АДРЕСОВ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ // Al-Farg'oniy avlodlari. 2024. №4. URL: https://cyberleninka.ru/article/n/obnaruzhenie-i-otsenka-fishingovyh-url-adresov-s-ispolzovaniem-algoritmov-mashinnogo-obucheniya (дата обращения: 22.04.2025).

ISSN 3030-3907