УДК 004

Ишмухаметов И.Р., Чернышев К.В.
Уфимский университет науки и технологий, Уфа

Научный руководитель:
Юнусова Д.С.
Уфимский университет науки и технологий, Уфа

ОБНАРУЖЕНИЕ ТРОЯНСКИХ ПРОГРАММ

Аннотация. Статья посвящена исследованию применения различных моделей машинного обучения для выявления троянских программ. Выполнен сравнительный анализ эффективности моделей Decision Tree, Random Forest, K Nearest Neighbors и Stacked Classifier с использованием метрик Accuracy, Precision, Recall, F1 Score и AUC-ROC. Основной задачей работы является определение наиболее точного и надежного метода для своевременного обнаружения и противодействия троянским программам.

Ключевые слова: троянская программа, машинное обучение, классификация, Decision Tree, Random Forest, K Nearest Neighbors, Stacked Classifier, Accuracy, Precision, Recall, F1 Score, AUC-ROC.

Троянские программы представляют собой вредоносное программное обеспечение, которое маскируется под легитимные приложения, предоставляя злоумышленникам доступ к персональным данным и ресурсам пользователя. Выявление таких программ важно, так как они представляют собой серьезную угрозу. Могут похищать конфиденциальные данные, такие как пароли и банковские реквизиты, используя для этого различные методы, включая перехват ввода с клавиатуры. Кроме того, они могут вызывать финансовые потери, проводя несанкционированные транзакции или создавая поддельные счета. Не менее опасно то, что некоторые троянские программы имеют способность повреждать систему, удаляя или изменяя файлы и тем самым делая компьютер непригодным для использования. Традиционные методы, основанные на поиске известных сигнатур вирусов, становятся менее эффективными, поскольку современные угрозы постоянно эволюционируют. Именно поэтому актуальным решением стало применение машинного обучения, способного выявлять скрытые паттерны поведения зараженных устройств.

Статья посвящена разработке эффективных методов идентификации троянских программ с использованием методов машинного обучения. Мы рассматривали набор данных, состоящий из 177 482 наблюдений, среди которых 90 683 соответствовали случаям заражения («Trojan») и 86 799 – чистым («Benign»). Данные представляли собой подробное описание технических аспектов сетевого взаимодействия устройств, таких как временные промежутки, объемы передаваемых данных, частоты передачи пакетов и другие статистические параметры.

Методика анализа состояла из нескольких последовательных шагов. Во-первых, произошла очистка данных от пропусков и формирование нового набора признаков, позволяющего точнее определить отличия между поведением зараженного и незараженного устройства. Затем осуществлялось преобразование категорических признаков в числовой формат, балансировка классов с помощью метода SMOTE и стандартизация данных для устранения влияния размаха величин.

Далее выполнялось сравнение нескольких популярных моделей машинного обучения: дерева решений, случайного леса, k ближайших соседей и ансамблевого классификатора. Чтобы выбрать лучшую модель, каждая из них прошла процедуру оптимизации гиперпараметров с целью максимизировать точность классификации.

В табл. 1 представлены итоговые результаты сравнения моделей, оцениваемые по основным показателям качества: точности, полноте, F1-мере и площади под кривой ROC (AUC-ROC).

Таблица 1

Показатели эффективности моделей машинного обучения при определении троянских программ

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Decision Tree	0,95	0,95	0,95	0,95	0,95
Random Forest	0,95	0,95	0,95	0,95	0,95
K Nearest Neighbors	0,91	0,91	0,89	0,90	0,91
Stacked Classifier	0,95	0,95	0,95	0,95	0,95

Полученные данные показывают значительное преимущество Decision Tree и Random Forest, которые достигли примерно одинаковых высоких уровней точности порядка 95 %, что свидетельствует о стабильной и точной работе этих моделей.

Модель K Nearest Neighbors, несмотря на чуть меньшую общую производительность, остается достаточно конкурентоспособной. Ансамблевый классификатор также подтвердил свою надежность, показав аналогичные результаты с деревьями решений и случайным лесом.

Результаты подчеркивают перспективность применения машинного обучения для детектирования троянских программ на основе анализа сетевого трафика. Новый подход позволил создать высокоэффективные инструменты, способные успешно выявлять угрозы в режиме реального времени, открывая возможности для совершенствования современных защитных механизмов и средств кибербезопасности.

Список использованных источников:

1. Кобяков Н.С. Оценка опасности вредоносных программ классов «Троянские программы» / Н.С. Кобяков // Альманах Пермского военного института войск национальной гвардии. 2023. № 2 (10). С. 31–38.

2. Фадеева М.А. Применение метрик качества моделей бинарной классификации для анализа модели данных / М.А. Фадеева // Инженерные кадры – будущее инновационной экономики России. 2023. № 1. С. 1104–1107.

Ishmukhametov I.R., Chernyshev K.V.
Ufa University of Science and Technology, Ufa, Russia

Scientific supervisor:
Yunusova D.S.
Ufa University of Science and Technology, Ufa, Russia

TROJAN PROGRAM DETECTION

Abstract. The article is devoted to the study of the application of various machine learning models to identify Trojans. A comparative analysis of the effectiveness of the Decision Tree, Random Forest, K Nearest Neighbors, and Stacked Classifier models was performed using Accuracy, Precision, Recall, F1 Score, and AUC-ROC metrics. The main objective of the work is to determine the most accurate and reliable method for timely detection and countering Trojan programs.

Keywords: trojan program, classification, Decision Tree, Random Forest, K Nearest Neighbors, Stacked Classifier, Accuracy, Precision, Recall, F1 Score, AUC-ROC.