ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

FRAUD APP DETECTION OF GOOGLE PLAYSTORE APPS USING MACHINE LEARNING

Mr. K. VIJAY KUMAR, Assistant Professor

SOHAIL AHMED, SHAIK SHOYEB, MD.KHIZAR UDDIN, SHAIK ALTHAF, T SAMPATH

SRI INDU COLLEGE OF ENGINEERING AND TECHNOLOGY

Sheriguda (V), Ibrahimpatnam (M), RangareddyDist – 501 510

ABSTRACT

Along the rise in the various mobile applications which are used in daily life, it's more necessary than ever to stay on top of things to decide which are safe and which don't. It is impossible to pass judgment. Our system is based on four parameter that include ratings, reviews, in app purchases and Contains ad to predict. Our system compares three models Decision Tree classifier, Logistic Regression and Naïve Bayes. These models were further analyzed on four parameters of F1 score, Recall, Precision and Accuracy. A good F1 score should be greater than 0.7 and a recall score greater than 0.5 is considered to be good with higher precision and accuracy. On analysis we found Decision tree model as a good model with accuracy of 85%, F1score of 0.815, Recall value of 0.85 and precision of 0.87

INDEX : mobile application, daily life, safe not safe, decision tree classifier, logistic regression.

1. INTRODUCTION

The Google Play Store stands as a vibrant marketplace hosting millions of mobile applications, catering to a diverse range of user needs. However, within this expansive ecosystem lies a persistent threat: fraudulent applications. These deceptive apps not only jeopardize users' financial security but also pose significant risks of data breaches and privacy violations. To address this growing concern, the application of machine learning (ML) techniques emerges as a crucial strategy.Traditional approaches to fraud detection often struggle to keep pace with the evolving tactics of malicious actors. Rule-based systems, while useful, may fail to detect subtle patterns indicative of fraudulent behavior. In contrast, machine learning offers a dynamic and adaptive approach, capable of discerning intricate patterns from vast datasets.

This paper delves into the realm of fraud detection within the Google Play Store, leveraging the power of machine learning algorithms. Our objective is to develop a robust framework capable of identifying and flagging potentially fraudulent applications with high precision.

By analyzing prevalent fraud patterns and tactics employed by malicious actors, we seek to gain insights into the evolving landscape of app-based fraud. Through empirical evaluation, we aim to assess the efficacy of various machine learning algorithms in detecting fraudulent activities, considering factors such as feature engineering, model selection, and performance evaluation metrics.Our ultimate goal is to contribute towards building a scalable and adaptable fraud detection system that can effectively counter emerging threats within the dynamic app ecosystem. By enhancing the security and integrity of the Google Play Store platform, we endeavor to create a safer and more trustworthy environment for both users and developers.

2. LiteratureSurvey

Title: Detection Of Fraud Ranking For Mobile App Using IP Address Recognition Technique

Abstract: Fraudulent activity in mobile app market means intruders or app developers use shady means to increase their app rating to bring their app in top 20 list to inflate their app sale. Knowledge engineering domain usually uses the methodologies to extract the useful knowledge from the givenlarge data. Ongoing rapid growths of online data have created the need of KDD. Also ongoingrapid growth of online rating and review system to the app, make fraud app has been launchedin the mobile market and let them be downloaded and used by many users. The fraud mobileapp is not worth to use and wasting device memory. Sometimes such app is created with malicious software which is harmful to the device. To avoid this situation the fraud app should be find out. Inexisting work, fraud rankings are detected by applying the mining algorithm in app review.Local anomaly was detected instead of global anomaly. The analysis had been done reported.Human evaluators evaluated and produce the result. Time complexity is more to evaluate.To overcomethis drawback, FRDS is proposed. To detect the fraud app, app's reviews should be checked. Tocheck whether the app reviews are fraud or not,the Fraud Ranking Detection System is proposed.In the mobile market, each mobile has its own unique IP Address. Hence, each user has uniqueIP Address. When giving the reviews to app, user IP Address is extracted by using the IPAddress recognition technique.So that, fromone IP Address numberof reviews cannot be providedto same app. In this way fraud review is prevented in proposed work. This approach decreasesthe evaluation time of the result, hence it isefficient than the existing approach.Index Terms DetectingFraud Ranking, Ratings and Reviews, Aggregation method.

2.Title: "**An enhanced mining leading session algorithm for fraud app detection in mobile applications**"

Abstract: Now days, mobile App is a very popular and well known concept due to the rapid advancement in the mobile technology and mobile devices. Due to the large number of mobile Apps, ranking fraud is the key challenge in front of the mobile App market. Ranking fraud refers to fraudulent or vulnerable activities which have a purpose of bumping up the Apps in the popularity list. In fact, it becomes more and more frequent for App developers to use tricky means, like increasing their Apps' sales or posting fake App ratings, to commit ranking fraud. While the importance and necessity of preventing ranking fraud has been widely recognized. After understanding the details of ranking fraud and the need of ranking fraud detection, the paper proposes a ranking fraud detection system for mobile Apps. The proposed system mines the active periods such as leading sessions of mobile apps to accurately locate the ranking fraud. These leading sessions can be useful for detecting the local anomaly instead of global anomaly of App rankings. Besides this, by modeling Apps ranking, rating and review behaviors using statistical hypotheses tests, we investigate three types of evidences, they are ranking based evidences, rating based evidences and review based evidences. Furthermore, we propose an aggregation method based on optimization to integrate all the evidences for fraud detection. Finally, the proposed system will be evaluated with real-world App data which is to be collected from the App Store for a long time period.

3.Title: "**A Methodology to Detect Fraud Apps Using Sentiment Analysis**"

Abstract: In Today's world, smart phone are very important in our daily life. In today's Scenario everyone is using smart phone. Nowadays, there are numerous applications out there on web due to that user cannot continuously get correct or true reviews concerning the merchandise on web. There are so many fraud applications on the internet.The growth of apps was increased by 1.6 million at App Store and Go ogle Play. There are many apps fromwhich any app can be fraud, so the identificatio n of true app is needed.Our fraud app detection application will help user to identify which application is true. Our main target is todetect fraud app because there are huge no of mobile apps. By analyzing admin declare app fraud app and alsobased on us er comment evidence we give rating.

4.Title: "Detection of fraud apps using sentiment analysis"

Rank misrepresentation in the portable Application advertise alludes to extortion/misleading exercises whose lone object is to have a reason for hitting up the Applications in the prominence

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

list. It turns out to be more incessant for Application designers to utilize terrible means, for example, expanding their Application deals or posting false App evaluations, to confer positioning extortion.It is vital to avoid positioning fraud as there is restricted comprehension and research in this field. Up till now, in this paper,we have given a comprehensive perspective of positioning misrepresentation and recommended positioning extortion identification framework.

5.Title: "A Survey of Sentiment Analysis techniques"

Sentiment analysis is an application of natural language processing. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. It helps in human decision making. To perform sentiment analysis, one has to perform various tasks like subjectivity detection, sentiment classification, aspect term extraction, feature extraction etc. This paper presents the survey of main approaches used for sentiment classification.

6.Title: "Detecting Spam Web pages through Topic and Semantics Analysis"

Web spam is an illegal and immoral way to increase the ranking of web pages by deceiving search engine algorithms. Therefore, different methods have been proposed to detect and improve the quality of results. Since a web page can be viewed from two aspects of the content and the link, the number of extracting features is high. Thus, selection of features with high separating ability can be considered as a pre processing step in order to decrease computational time and cost. In this study, a new backward elimination approach is proposed for feature selection. The main idea of this method is measuring the impact of eliminating a set of features on the performance of a classifier instead of a single feature which is similar to the sequential backward selection. This method seeks for the largest feature subset that their omission from whole set features not only reduces the efficiency of the classifier but also improves it. Implementations on WEBSPAM-UK2007 dataset with Naive Bayes classifier show that the proposed method selects fewer features in comparison with other methods and improves the performance of the classifier in the IBA index about 7%.

7.Title: "In Data Mining (ICDM)"

The IEEE International Conference on Data Mining (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

and practical development experiences. The conference covers all aspects of data mining, including algorithms, software, systems,and applications. ICDM draws researchers, application developers, and practitioners from a wide rangeof data mining related areas such as big data, deep learning, pattern recognition, statistical and machine learning, databases, data warehousing, data visualization, knowledge-based systems, and high-performance computing. By promoting novel, high-quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to advance the state-of-the-art in data mining.

3. PROBLEM STATEMENT

In the context of the increasing prevalence of mobile applications in daily life, the need to discern their safety has become more crucial than ever. Passing judgment on each application individually is impractical, prompting the development of a systematic approach. The proposed system relies on four key parameters—ratings, reviews, in-app purchases, and the presence of ads—to make predictions about the safety of mobile applications. A comparative analysis was conducted using three machine learning models: Decision Tree classifier, Logistic Regression, and Naïve Bayes. The evaluation focused on four performance metrics—F1 score, Recall, Precision, and Accuracy. Typically, a good F1 score is considered to be greater than 0.7, while a recall score exceeding 0.5 is deemed satisfactory, along with higher precision and accuracy. The analysis revealed that the Decision Tree model emerged as a robust performer, boasting an accuracy rate of 85%, an F1 score of 0.815, a Recall value of 0.85, and a precision of 0.87. These results position the Decision Tree model as a favorable choice for predicting the safety of mobile applications in this system.

Existing System Disadvantages:

Subjective Interpretation: The reliance on customer ratings and reviews introduces subjectivity into the assessment process. Different users may interpret and rate an app differently, making it challenging to ascertain genuine feedback from fraudulent ones.

Limited Scope: The existing system's focus on a narrow set of parameters overlooks other potential indicators of fraudulent behavior. Malicious apps can employ tactics that bypass the current detection criteria.

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

Inefficient Manual Assessment: Relying on manual assessment of each app's authenticity is impractical due to the vast number of applications available. This inefficiency hinders the ability to effectively identify and mitigate fraudulent apps.

4. PROPOSED SYSTEM

We propose a system that would identify such fake applications on the play or app store. We can acquire the probability of determining whether an app is fake or not, therefore we present a system that uses four features that are in app purchases, contains ad ,ratings and reviews to determine the probability of an app whether it's scamming its consumers or not.Nevon projects has proposed such a comprehensive framework that can be expanded with additional evidence generated by the domain to detect quality fraud. It is one of the most advanced projects to detect fraudulent applications using information algorithms. This program provides only 75-80% accuracy in detecting fraudulent applications.The sole purpose of the given proposed system is majorly to review the fraud detection of google play store applications and to use the four parameter methods to differentiate certain fraudulent applications or commonly referred to as spam applications. Experimental analysis is performed on different types of methodology in the proposed manner for the detection of fraud or fake applications. Our system will receive fraud with four types of evidence, such as ad-based ratings, in-app purchases and evidence-based reviews. In addition, the development-based integration approach incorporates all four aspects to detect fraud. Various machine learning model were implemented which provided different results for the accuracy. By analysis, we found that our given proposed method provides 85% accuracy compared to other algorithms. While independent thinking still exists, the decision tree section performs better compared to other models such as the recession and the naïve bayes. It is an intuitive algorithm for separation problems. It is a reliable real-time guess, a setback problem. Decision trees can manage non-linear data sets effectively. It plays a role in decision-making in various fields of life, including engineering, social planning, business, and even law. In this pursuit, we implemented various machine learning models, each yielding different accuracy results. Our thorough analysis revealed that our proposed method achieves an accuracy of 85%, outperforming alternative algorithms. Among these models, the decision tree algorithm stood out, demonstrating superior performance compared to other approaches like regression and naïve Bayes. The decision tree algorithm excels in solving complex separation problems, providing reliable real-time predictions, and efficiently handling non-linear datasets. Its applicability extends across diverse domains, including engineering, social planning, business, and legal contexts. Our system, with its comprehensive approach and utilization of advanced

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

machine learning techniques, represents a significant step forward in combating fraudulent applications. Its robust accuracy and adaptable framework contribute to the ongoing efforts to ensure the integrity and security of app marketplaces.

Advantages of Proposed System

The proposed system offers a range of advantages, including robust fraud detection, automated analysis, real-time decision-making user protection and a commitment to transparency. By addressing the challenges of fraudulent applications head-on, the system contributes to the overall integrity and security of app marketplaces.

5. IMPLEMENTATION

Data Collection

Collecting dataset of 5500 apps from Kaggle and fraud apps from various articles.Kaggle is a tremendous platform in which it is easy for every user to discover and post datasets, check and construct models with other information. Scientists and Machine learning engineers solve data challenges that require information via the usage of statistics from Kaggle. Kaggle provides a no-setup, custom-made, Jupyter Notebooks environment. It also provides access free to GPUs and a massive repository of various community published data & code.

Data Pre-processing and filtration

Remove missing value, only required columns are kept. This is done through pandas. Pandas is used especially for facts evaluation. Panda imports statistics from diverse document formats including comma-separated values, JSON, square, and MS Excel. Pandas allows for a filtration of facts together with merging, reshaping, sorting, and data processing, in addition to facts conflicts.

Model creation and training

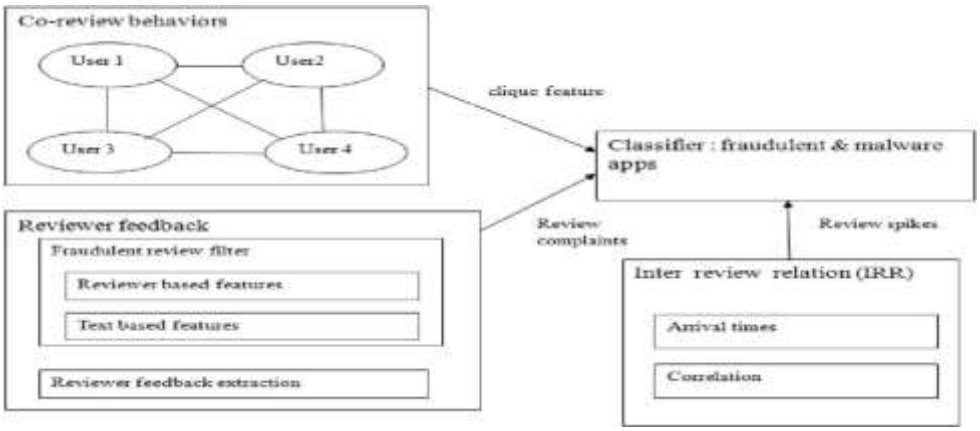
In the model we have taken 75% of the data for training and rest 25% for testing.We have Made model of decision tree, naive bayes algorithm and logistics regression. Further after comparing all the models, the accuracy of decision tree was highest.Compared accuracy of Naive Bayes (83%), Logistic Regression (84%) and Decision Tree (85%) algorithm.

Deploying of model

Made user interface web page using flask. In which the user puts the url in the search box and gets the result as "fake" or "safe to use" for the app. Flask is a web framework, because of this flask offers you with gear, libraries and technology that allow you to build a web utility.

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

6. SYSTEMARCHITECTURE



7. SCREENSHOTS

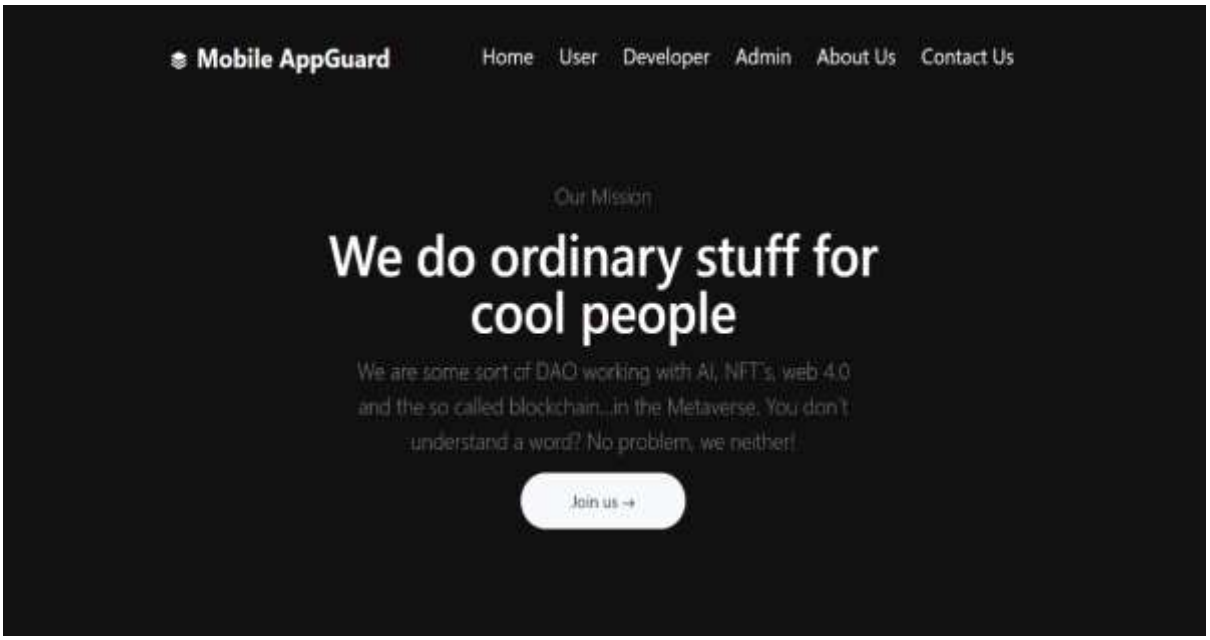


Fig Home Page

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

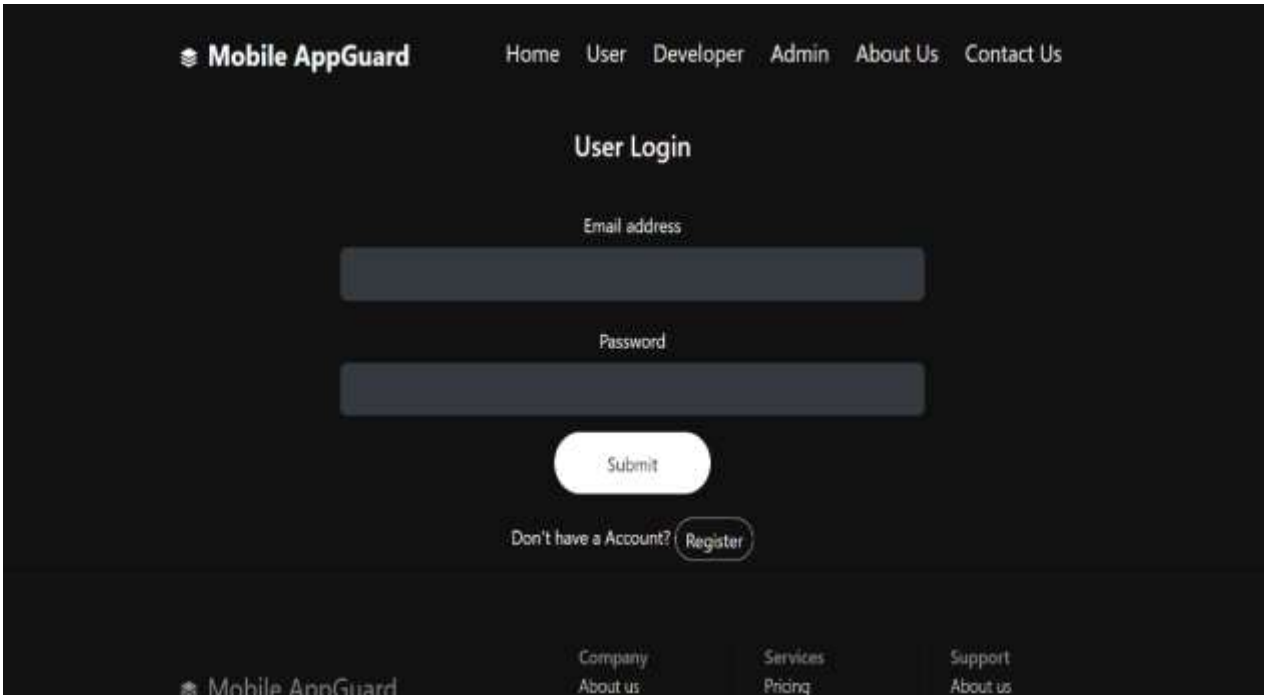


Fig User Login Page

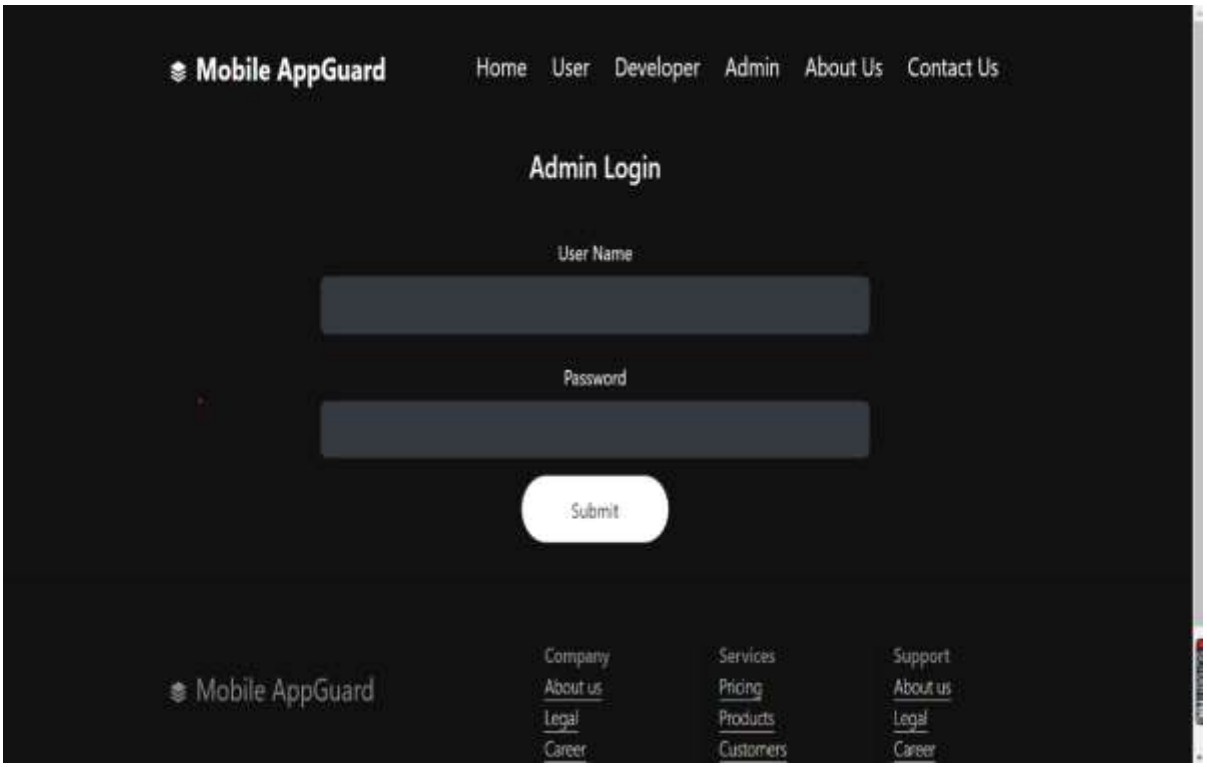


Fig Admin Login Page

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

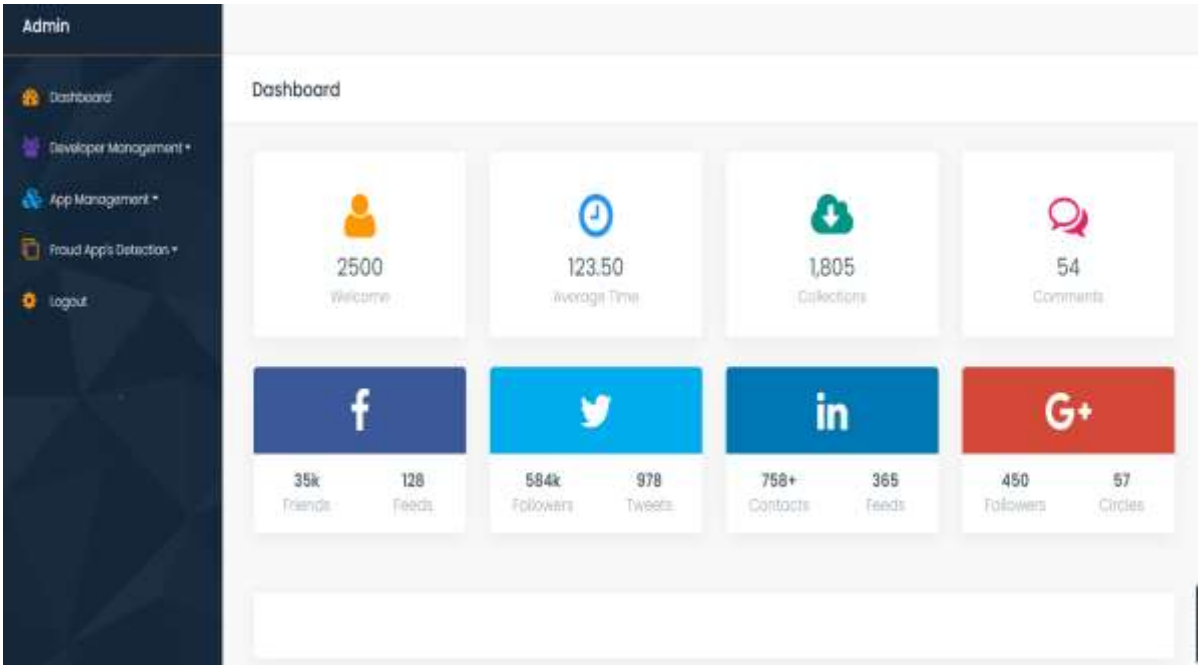


Fig Admin Dashboard

UPLOAD :On this page, the user of the system can upload .csv file. The user has to select the file by clicking on the Choose file button

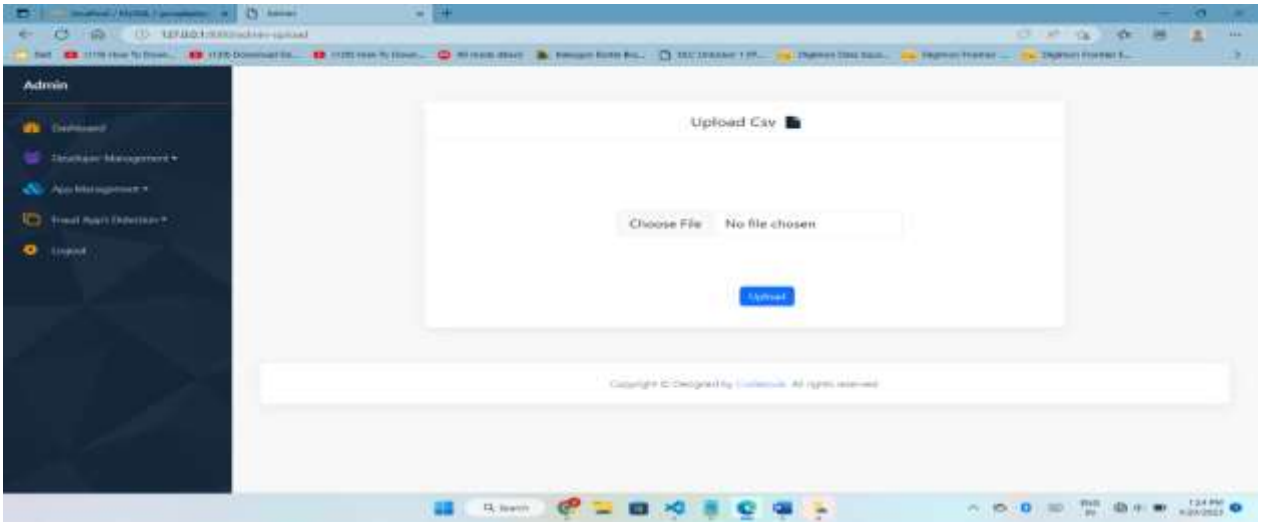


Fig Admin Upload Dataset

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

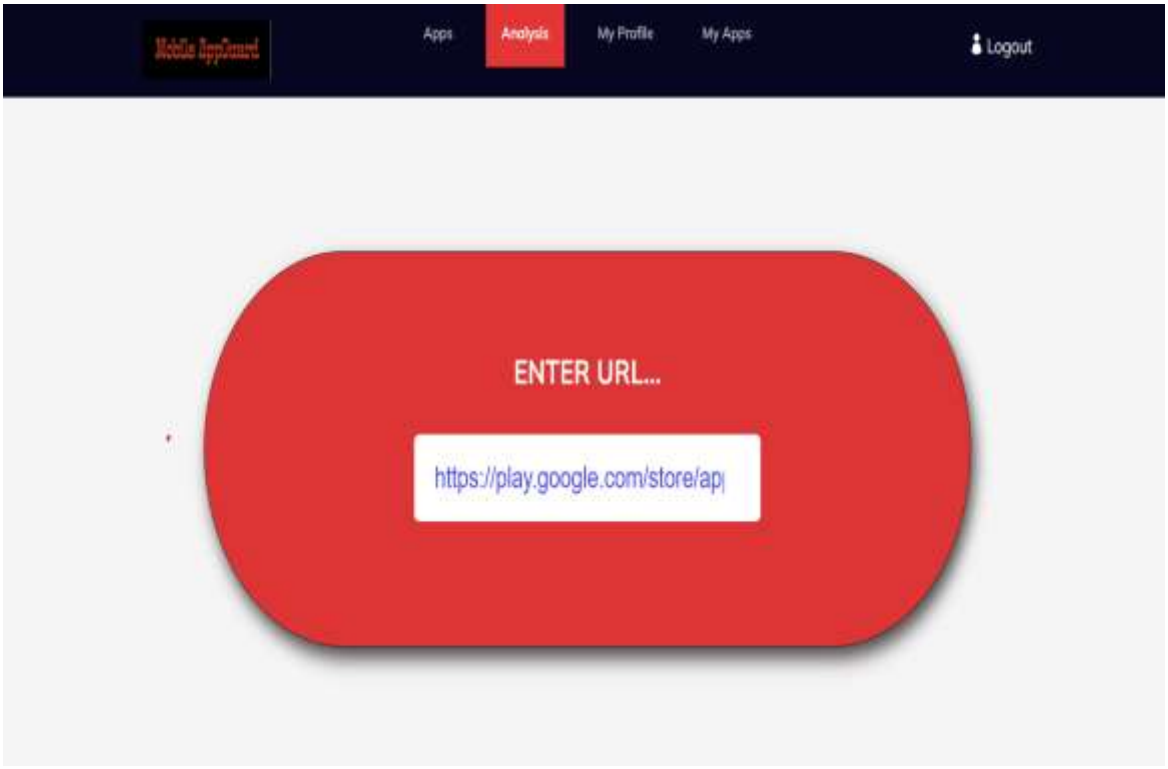


Fig Link Upload Page

Result

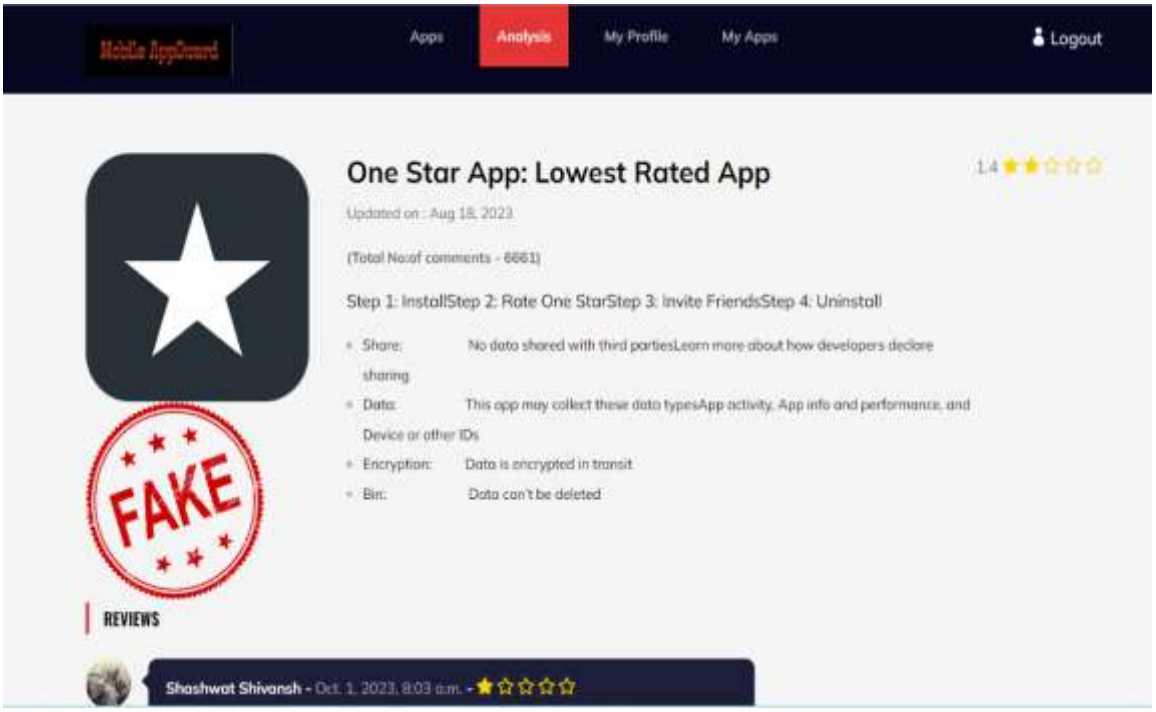


Fig Negative Result

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

1. 

Fig Positive Result

8. .CONCLUSION

In today's rapidly advancing technological landscape, the issue of security has emerged as a critical concern. The proliferation of digital platforms has brought about unprecedented conveniences, but it has also given rise to significant security threats. One such challenge is the proliferation of fraudulent applications within Google's app stores. These malicious apps not only jeopardize users' personal privacy and sensitive data but also compromise the overall digital ecosystem's integrity.

In response to this pressing concern, our research endeavors have been meticulously directed towards developing a robust solution for the detection of fraudulent software. We have focused on a comprehensive approach, incorporating four key parameters: app scales, review scores, in-app purchases, and content additions. By analyzing these crucial facets, we aim to effectively distinguish between legitimate applications and those that pose a threat to users.

The cornerstone of our approach lies in the application of advanced algorithms to assess the authenticity of apps. To this end, we conducted a meticulous comparison of three prominent algorithms: Decision Tree, Naive Bayes, and Logistic Regression. Remarkably, our analysis

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

revealed that the Decision Tree algorithm demonstrated the highest accuracy rate, achieving an impressive 85%.

Our proposed framework boasts inherent measurability and scalability, positioning it as a versatile tool for fraud detection across a variety of domains. Its modular design allows for the incorporation of additional domain-specific evidence, further enhancing its efficacy in identifying fraudulent activities. This extensibility underscores the adaptability of our system to future challenges and emerging fraud tactics.

Through rigorous experimentation and empirical analysis, we have successfully demonstrated the potency of our proposed system. The accuracy of the algorithmic detection mechanism, coupled with its measured performance, showcases its potential to serve as a critical tool in combating fraudulent operations. Moreover, our system contributes to standardizing fraud detection operations, streamlining the process of distinguishing between legitimate and malicious applications.

Ultimately, our innovative approach holds promising implications for enhancing app store security. By effectively identifying and rating fraudulent applications, our system contributes to the overall integrity of digital marketplaces. This, in turn, bolsters user confidence, promotes privacy, and safeguards sensitive data. As technology continues to evolve, our research serves as a testament to our commitment to staying at the forefront of security solutions and protecting users in an increasingly interconnected digital world.

9. Future Enhancement

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

In the realm of app security, enhancing our fraud app detection project holds exciting prospects. Beyond Decision trees, advanced models like Random Forest and deep learning could bolster accuracy. Deeper feature engineering, incorporating NLP for app descriptions and analyzing user interactions could yield richer insights. Real-time analysis, transparent AI decisions, and continuous learning are vital for adaptive systems. Integrating user feedback, geolocation data, and cross-platform detection can expand scope. Regulatory and collaboration with experts ensure ethical and effective development. By exploring blockchain and embracing smart contracts, you can fortify app verification. As technology advances, our project has agility and innovation will be essential to combating emerging fraud tactics.

10. REFERENCES

- Esther Nowroji, Vanitha, "Detection Of Fraud Ranking For Mobile App Using IP Address Recognition Technique", vol. 4.
- Prasad, D. C. G. V. N., Bhargavram, K., &Guptha, K. G. (2015). Challenging Security Issues of Mobile Cloud S.R.Srividhya, S.Sangeetha – "A Methodology to Detect Fraud Apps Using Sentiment Analysis"
- Keerthana. B, Sivashankari.K and ShaisthaTabasum.S, "Detecting Malwaresand Search Rank Fraud in Google Search Using Rabin Karp Algorithm", IJARSE,7(02), 2018, pp.504-527.
- Nagamani, K., Prasad, C. G., &Chatrapati, K. S. (2016, February). A novel framework for optimal component based data center architecture. In 2016 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 1-10). IEEE.
- HarpreetKaur, VeenuMangat and Nidhi, — "A Survey of Sentiment Analysis techniques"
- International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017, pp.921
- Narasimha Chary Ch, CH. GVN Prasad. Human Deep Skin Surface Vibration Frequency Detection from CT and DT Signals Using Genetic Algorithms. International Journal of Algorithms Design and Analysis Review. 2024; 2(1): 1–8p.
- Navdeep Singh, Prashant Kr. Pandey and Mr.Srinivasan, — "Improved Discovery of Rating Fake for Cellular Apps", IEEE International Conference on Science Technology Engineering and Management (ICONSTEM), 2016, pp. 135-140.
- Weiman Wang, Restricted Boltzmann Machine. GitHub. Aug 2017. [Online] Available: https://github.com/aaxwaz/Fraud-detection-usingdeep-learning/blob/master/rbm/rbm.py.

ISSN: 2278-4632
Vol-14, Issue-7, June: 2024

- CHOLLETI, N., & HIRWARKAR, T. (2020). BIOMEDICAL DATA ANALYSIS IN PREDICTING AND IDENTIFICATION CANCER DISEASE USING DUO-MINING. Advances in Mathematics: Scientific Journal, 9, 3487-3495.
- Ranking fraud Mining personal context-aware preferences for mobile users. H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, and J. Tian. In Data Mining (ICDM), 2012 IEEE 12th International Conference on, pages1212–1217, 2012.
- R. Shobarani, R. Sharmila, M. N. Kathiravan, A. A. Pandian, C. Narasimha Chary and K. Vigneshwaran, "Melanoma Malignancy Prognosis Using Deep Transfer Learning," *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, Bangalore, India, 2023, pp. 1-6, doi: 10.1109/ICAIA57370.2023.10169528
- Detecting product review spammers using rating behaviors. E.-P. Lim, V.-A.Nguyen, N. Jindal, B. Liu, and H. W. LauwIn Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM ‟10 pages 939–948, 2013.
- Detection for mobile apps H. Zhu, H. Xiong, Y. Ge , and E. Chen. A holistic view. In Proceedings of the 22nd ACM international conference on Information and knowledge management, CIKM ‟13