A study on techniques to combat fraud and price variations in E-Commerce Applications using Blockchain and Machine Learning

Dr. V. Baby [1], G. Avinash Reddy[2], K. Yogendra [3], K. Sree Harshitha[4], N. Padmini Chowdary[5]

[1]*Associate Professor, Dept. of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

[2, 3, 4, 5] *Student, Dept. of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

---***---

Abstract - *E-Commerce applications have revolutionized the way customers make purchases. Generally, if a user wants to buy any product, first the user should go to a particular retailer, then select the products to buy and make a payment. This is a long and tiresome process. Hence, users are looking for a simple digitized solution that will make the process of buying products easier. This digitized solution of buying and selling goods online is described as E-Commerce. But this approach has created some problems. For example, the products delivered to the people may be fake or/and the vendors who sold these items are fraudulent. Another issue is that the reviews of some products might not be correct, very similar products can be sold at varying prices. The quality of the products may not be good. Also, there are a large number of items that are present and customers might find difficulties in finding the exact item that they are looking for or an item they truly desire. This paper mainly deals with various existing methodologies and drawbacks related to the above problems like Fake Product Detection, Product Price Suggestion (Prediction), Product Quality Detection from Reviews, Spam Review Detection, and Fake Seller Detection.*

Key Words: **Blockchain, Price Suggestion, NLP, Fake Products, Fraudulent Sellers, Spam Reviews, Product Quality Detection**

1. INTRODUCTION

Electronic commerce, or e-commerce, is the exchange of products and services over a network, such as the internet. If a user wants to buy the goods offline then the products are real and tangible and the user can touch and feel the product and assess the quality before buying it. Still, in the current digital age, everyone wants to buy products online where the products are represented virtually in the form of images and videos. Hence there are a lot of problems that users are facing such as users are not sure whether the product which is being sold is real or fake, they are not able to determine the quality of the product, they don't know about the right price of the product and the website may also contain spam reviews to the product.

When a user wishes to purchase a product through an online store, the user reviews the product specifications before placing an order, however the consumer might receive a counterfeit goods. For example, the user might have placed an order for PUMA shoes but the shoes that are delivered look exactly as the PUMA shoes but they are not. Such experiences make users lose faith in the brand because the quality of the product delivered by the fake brand is not as good as that of the real product. So genuine brands lose their customers without their mistakes. So, this problem of selling fake products has a negative effect on both the customers and manufacturers.

Users can also give reviews about the product that they have brought so that other users can read them and decide whether or not to buy the product. Hence manufacturers of the product can create a lot of fake accounts just to give positive reviews to the products to increase sales and eventually if the product is not good then users may fall into the trap of buying a low-quality product. Similarly, competitors can create many fake accounts just to give a lot of negative reviews about the product to decrease their sales.

The manufacturers have the freedom to determine the price of the products, this gives the manufacturers the leverage to sell the products at a very high price whereas the competitors are selling the same product at a lower price. The users may fall into this trap by paying large amounts to buy a product only to find the same product at a lower price.

These are the problems that are being faced by both the manufacturers and the users. Users lose their trust in the brands and manufacturers will lose the customers and both of them will face financial losses. There should be mechanisms to tackle such problems, otherwise, people might stop buying the products online.

2. RELATED WORK

Wasnik, K. [1] The manufacturer adds the product details to the database and generates a QR code, and sends it with the product to the supplier. Next, the supplier scans the QR and adds his details, and sends this product to the customer. Finally, the customer scans the QR to check the authenticity of the product. They have used ReactJS for the frontend, the blockchain that is used is Ethereum, web3.js is used for connecting the javascript code with the Ethereum blockchain, metamask valet is used to manage the accounts, and MySQL database to store the data.

Shreekumar, T. [2] Developed a mobile application, Used dart for the frontend, Node.js for writing the server-side code, and the database used is Firebase and they have used Ethereum-based blockchain and solidity for writing the smart contracts, web3.js is used to connect the javascript code with the blockchain. They have used the QR-Code-based system to determine the authenticity of the products. A notification verifying the legitimacy of the goods will be given to the consumer if the QR Code matches; else, a notification stating that the product is false or counterfeit will be issued.

Jambhulkar, S. [3] Various users involved are Manufacturers, Distributors, Retailers, and Customers. The manufacturer will add the data to the blockchain, distributor and retailer will update the data in the blockchain by adding their information. Whenever the user scans the QR Code, they will be able to access all the information, from the manufacturer to the retailer. The front end is developed using HTML, and CSS and they have used PHP for server-side code, and the database used is MySQL.

Ma, J., Lin [4] Developed smart contracts using solidity. The data is stored in Geth, which is a private Ethereum-based blockchain and can be accessed using Meta mask. Manufacturers will include information about themselves and the seller, such as the seller's address and the quantity of products that can be sold from a certain seller. Then the product is sent to the supplier by attaching a QR code to it. The supplier will send this product to a specific user who has bought it. This application also provides the feature of exchanging the product where the manufacturer initially verifies the identity of the user and then the exchange process will be initiated by changing the status in the blockchain.

Robin, Md. Rakibul Hassan [5] Used HTML, CSS, and JavaScript for the frontend, PHP for the server side and the database which is used is MySQL. Used Ganache which is a local Ethereum-based blockchain and meta mask account for storing Ethereum tokens. The manufacturer adds the product details and address, the distributor and retailer will update the details by adding their addresses, and finally, the user will be able to view the details of the manufacturer, distributor, and retailer.

Tejaswini Tambe [6] Developed a mobile application. They used firebase as their database to store the data, Used SHA 256 algorithm to generate the unique QR code for each product, and these QR codes are scanned using the built-in feature present in the mobile application which they have developed without using a separate application only for the purpose of scanning the QR codes. When the QR codes are scanned it will provide a label indicating “fake product” if the product is fake, or it will be labelled as “received” if the product is not fake.

Prasad, A. K [7] They have developed a model to predict the price based on product images, and metadata such as composition, etc. With PriceNet as the baseline model, they first built a model to predict prices based on images only (VGG16 and InceptionResNetv2), then built a model to include metadata (NLP, XGBoost & LightGBM). The Two Stage XGBoost model achieved a best R^2 score of 0.78 compared to a baseline of 0.18 for VGG16 and 0.22 for InceptionResNetv2.

Zehtab-Salmasi, A. [8] They proposed several forecasting models to predict the price range of cell phones, based on their specifications. Based on the price range threshold, classes are assigned. Out of several models developed, CNN clubbed with Inception-V3 with Convolutional Image Feature Extraction and Dense Concatenation gave the best results, with an F1-Score of 88% on the CD18 dataset.

Mahoto, N. A. [9] This work proposed a method, to decide the price of the product which the seller should sell for, based on the customer’s buying behaviour, which is a Business Intelligence-Machine Learning fusion model. They built three models: Random Forest, Logistic Regression, and One-vs-All model. The One-vs-All model had a mean Recall and Precision of around 90%.

Han, L. [10] Proposed a model that provided price suggestions for non-identical items across different categories. First, based on product images, it is classified as qualified or not. If found not qualified, the seller will have to give a text description. Finally, based on the photographs and text description, a pricing recommendation is made for the product. RMSLE (Root Mean Squared Logarithmic Error) for the project was a respectable 0.69.

Fathalla, A. [11] For estimating the price of a used item based on the picture and written description of the item for various item kinds, a deep model architecture composed of LSTM & CNN was proposed. The model obtained a respectable R^2 score of 0.77.

Katarya, R [12] This paper proposed CapsMF, which is a combination of Capsule Network and Matrix Factorization. Capsule network uses embedding layer and Bi-Directional Gated recurrent Unit for representation of textual descriptions of users and items and the Matrix Factorization is used to generate the improved recommendations. The cold start problem with recommender systems was addressed by this approach.

Hwangbo, H [13] This paper developed a model called K-recSys, which is an extension to the existing collaborative filtering recommendation system. It considered both online and offline preferences of the users to generate the recommendations. When this model is tested in the actual operating environment then it is found that the model adopted substitute recommendations more frequently than the complementary recommendations.

Hendrawan, R [14] This approach proposes a method to assess the quality of reviews as they are helpful to the user in making a buying decision. The quality of the review is determined based on 3 characteristics: structural, readability and metadata. The weighted sum is calculated for the reviews and they are sorted by considering the final score in the order of highest to lowest.

Singla, Z., [15] In this paper the sentiment of reviews are analyzed as they provide insights about the performance of the products and customer satisfaction. The dataset is collected from Amazon and an under sampling method is used to deal with the data imbalance problem. They have trained various classification models like Naive Bayes, Decision trees, and SVM. The accuracy of the models is identified using 10-fold cross-validation. The highest accuracy is achieved for SVM which is 81.75% and Naive Bayes has the worst accuracy.

Song, J., Qu, X., [16] Proposed a model for collective fraud detection called subGNN, which is a subgraph based method. Initially, the subgraphs are extracted from the user behaviour. Then, they remove any entity related information, so that the model is entity-independent. Then they extracted the knowledge reasoning rules from heterogeneous subgraphs using a relational graph isomorphism network (R-GIN). The model is tested with publicly available datasets from Amazon and Yelp. When the SubGNN is used to detect fraudulent transactions on a newly collected dataset, it predicts the fraud ones with accuracy of 0.99 and more than 90% of fraud samples are recalled.

Dekou, R., [17] This paper proposed a model that collects, processes, and predicts the likelihood of a seller's listing data to be fraudulent, using existing machine learning libraries H2O AutoML and Catboost. They compared single machine learning models like Xgboost, Extremely Randomised Trees, Artificial Neural networks, and Generalised Linear Models with their stacked ensembles using validation sets. AutoML, which is a stacked ensemble model, provided the best performance with an F1 score of 0.73 when compared to other models.

Renjith, S. [18] This paper suggests a framework to detect fraud sellers in e-commerce platforms using machine learning. SVM is used for fraud detection, the model also uses the historical data in the marketplace to detect fraud sellers, so that a previously fraud seller reappearing with different details is still detected. The model is used to classify sellers as either fraud or not a fraud. The proposed model also considers inputs from fraud experts and they contribute to the refinement of the rules engine.

Maranzato, R., [19] This study is aimed at addressing the issue where few sellers try to deceive the reputation systems in e-commerce for their benefit. This approach describes certain characteristics in transactions that indicate fraud. Then they incorporated some other possible fraud characteristics and logistic regression is applied to both the datasets. The improved set containing other potential fraud characteristics performed better than the unimproved set with logistic regression, it increased the number of fraudsters identified by 110%.

Hooi, B., [20] This method primarily focuses on identifying users' or products' suspicious behaviour when it deviates from other accepted practises by building a Bayesian model for rating behaviour and developing a measure for spam identification. This method uses two alternative approaches: one where a single product receives multiple reviews from various users of the same text, and the other where a single person provides multiple reviews on other goods with identical reviews. When the model is run on the Flipkart dataset, it successfully detects fraud reviews in significant real-world applications with a precision of 84%. When this was applied to 250 of the application's most suspect users of Flipkart, 211 of them actually submitted reviews of fraud.

Liu, Y., [21] In this study in order to extract and analyse sentence representation through word embeddings, this method developed a hierarchical attention network. They used n-gram CNN to extract the multi-granularity and informative sentence representation, and they used BI-LSTM and convolution structures to extract complete data as well as the history of sentences. The model starts by comprehending the review's texture. On some datasets, this method's detection accuracy (F1) increased by 5% when it was evaluated on mixed and cross-domain datasets.

Liu, W., [22] This study analyses the review records and the results of the shopping data. To find outlier reviews, the isolation forest method is used to analyse the review data based on temporal factors. To demonstrate their method's usefulness and efficiency, they compared it to a number of other temporal outlier detection techniques. The suggested strategy can more effectively identify products that deviate from abnormally changing trends.

Shahariar, G., [23] This paper approaches by performing both traditional machine learning algorithms as well as deep learning techniques and compared both the performances and concluded that deep learning techniques were better. Here they also used three different deep learning classifiers like Multilayer perceptron, RNN(LSTM) and CNN .They used both labelled and unlabelled datasets preprocessed them using NLP methods and by active learning algorithm converted the unlabelled to labelled data and for feature selection TF - IDF for MLP and Word Embeddings for both CNN and LSTM. For the classification all the three classifiers performed well with high accuracies as compared to other existing techniques and among the three LSTM has higher accuracy.

3. CONCLUSIONS

With all the problems that E-Commerce applications have, it is difficult for a buyer to trust these systems anymore. We performed a literature survey on the related works that address these challenges. From the above literature survey, we can see that fake product detection is implemented using blockchain, but it does not prevent a fake manufacturer from registering on the blockchain as a genuine manufacturer. For predicting the price of a product, most researchers have made use of the product image, and the product description. However, images can sometimes be misleading and it can be hard to train a model which properly analyzes all the different categories of products. The existing research models for price prediction often did not make use of big datasets consisting of a variety of categories. For fake seller detection, the researchers used support vector machines, Logistic Regression, and other classification algorithms, considering features like product details, product listing accuracy, transaction details, product returns, the reason for returns, customer feedback, and complaints. However, the above solution can detect fraud when the fraudster reaches the peak, not in the initial stages. Some researchers also proposed graph-based models for the detection of collective fraud, but that address only one scenario of fraud. Most of the research related to fraud detection is often done by considering online auction systems and not marketplace applications. For spam review detection, the researchers used Machine Learning and Deep Learning techniques like the n-gram CNN method, Bayesian inference, and isolation forest method for identification and classification of the reviews as spam and non-spam.

ACKNOWLEDGEMENT

Special thanks to our team guide, Dr. V. Baby, for all the technical support and guidance which led to the completion of the literature survey phase of the project with a good result in the end.

REFERENCES

[1] Wasnik, K., Sondawle, I., Wani, R., & Pulgam, N. (2022). Detection of Counterfeit Products using Blockchain. In ITM Web of Conferences (Vol. 44, p. 03015). EDP Sciences.

[2] Shreekumar, T., Mittal, P., Sharma, S., Kamath, R. N., Rajesh, S., & Ganapathy, B. N. (2022). Fake Product Detection Using Blockchain Technology. JOURNAL OF ALGEBRAIC STATISTICS, 13(3), 2815-2821.

[3] Jambhulkar, S., Bhoyar, H., Dhore, S., Bidkar, A., & Desai, P. (2021). BLOCKCHAIN BASED FAKE PRODUCT IDENTIFICATION SYSTEM. International Research Journal of Modernization in Engineering Technology and Science, 2582-5208.

[4] Ma, J., Lin, S. Y., Chen, X., Sun, H. M., Chen, Y. C., & Wang, H. (2020). A blockchain-based application system for product anti-counterfeiting. IEEE Access, 8, 77642-77652.

[5] Robin, Md. Rakibul Hassan. (2021). Product Authentication Using Blockchain.

[6] Tambe, T., Chitalkar, S., Khurud, M., Varpe, M., & Raut, S. Y. (2021). Fake product detection using blockchain technology. International Journal of Advance Research, Ideas and INNOVATIONS in Technology, 7, 314-319.

[7] PRASAD, A. K., LARSON, M., & HIEMSTRA, D. (2021). Product Price Prediction.

[8] Zehtab-Salmasi, A., Feizi-Derakhshi, A. R., Nikzad-Khasmakhi, N., Asgari-Chenaghlu, M., & Nabipour, S. (2021). Multimodal price prediction. Annals of Data Science, 1-17.

[9] Mahoto, N. A., Iftikhar, R., Shaikh, A., Asiri, Y., Alghamdi, A., & Rajab, K. (2021). An intelligent business model for product price prediction using machine learning approach. Intelligent Automation & Soft Computing, 29(3), 147-159.

[10] Han, L., Yin, Z., Xia, Z., Guo, L., Tang, M., & Jin, R. (2021, January). Price Suggestion for Online Second-hand Items. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 5920-5927). IEEE.

[11] Fathalla, A., Salah, A., Li, K., Li, K., & Francesco, P. (2020). Deep end-to-end learning for price prediction of second-hand items. Knowledge and Information Systems, 62(12), 4541-4568.

[12] Katarya, R., & Arora, Y. (2020). Capsmf: a novel product recommender system using deep learning based text analysis model. Multimedia Tools and Applications, 79(47), 35927-35948.

[13] Hwangbo, H., Kim, Y. S., & Cha, K. J. (2018). Recommendation system development for fashion retail e-commerce. Electronic Commerce Research and Applications, 28, 94-101.

[14] Hendrawan, R. A., Suryani, E., & Oktavia, R. (2017). Evaluation of e-commerce product reviews based on structural, metadata, and readability characteristics. Procedia Computer Science, 124, 280-286.

[15] Singla, Z., Randhawa, S., & Jain, S. (2017, June). Sentiment analysis of customer product reviews using machine learning. In 2017 international conference on intelligent computing and control (I2C2) (pp. 1-5). IEEE.

[16] Song, J., Qu, X., Hu, Z., Li, Z., Gao, J., & Zhang, J. (2021). A subgraph-based knowledge reasoning method for collective fraud detection in E-commerce. Neurocomputing, 461, 587-597.

[17] Dekou, R., Savo, S., Kufeld, S., Francesca, D., & Kawase, R. (2021). Machine Learning Methods for Detecting Fraud in Online Marketplaces.

[18] Renjith, S. (2018). Detection of fraudulent sellers in online marketplaces using support vector machine approach. arXiv preprint arXiv:1805.00464.

[19] Maranzato, R., Pereira, A., do Lago, A. P., & Neubert, M. (2010, March). Fraud detection in reputation systems in e-markets using logistic regression. In Proceedings of the 2010 ACM symposium on applied computing (pp. 1454-1455).

[20] Hooi, B., Shah, N., Beutel, A., Günnemann, S., Akoglu, L., Kumar, M., ... & Faloutsos, C. (2016, June). Birdnest: Bayesian inference for ratings-fraud detection. In Proceedings of the 2016 SIAM International Conference on Data Mining (pp. 495-503). Society for Industrial and Applied Mathematics.

[21] Liu, Y., Wang, L., Shi, T., & Li, J. (2022). Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. Information Systems, 103, 101865.

[22] Liu, W., He, J., Han, S., Cai, F., Yang, Z., & Zhu, N. (2019). A method for the detection of fake reviews based on temporal features of reviews and comments. IEEE Engineering Management Review, 47(4), 67-79.

[23] Shahariar, G. M., Biswas, S., Omar, F., Shah, F. M., & Hassan, S. B. (2019, October). Spam review detection using deep learning. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0027-0033). IEEE.